Integration of 'omics data dissects genetic and metabolic causes of Macular Telangiectasia Type II

A thesis submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

By

Roberto Bonelli

ORCID: 0000-0003-2676-1230

Department of Medical Biology, The University of Melbourne

Population Health and Immunity Division

The Walter and Eliza Hall Institute of Medical Research

August 2019

Abstract

Macular telangiectasia type 2 (MacTel), is a rare degenerative eye disease. In this thesis, we will explore and analyse MacTel 'omics data including genomics, metabolomics and phenomics data to discover MacTel causal drivers that might be targeted for therapeutic interventions. The first chapter introduces the disease, the concept of data integration, and different types of 'omics data. The second chapter presents the first MacTel GWA study which discovered five loci. Four of them indicated a possible involvement of the glycine/serine metabolic pathway on the disease while the last pointed to genetic influences from retinal vasculature calibre. In the same study, glycine, serine and threonine were externally validated as metabolic biomarkers for MacTel. The third chapter presents a causal model for MacTel via genetic variants. Using Mendelian Randomization, we tested the genetic contribution to the disease of more than 140 metabolites and comorbid traits and found causal effects from serine and glycine, whilst eliminating other spurious metabolites like threonine. Serine was also found to cause disease progression based on longitudinal retinal image data for 455 patients. Novel genetic loci were further revealed via conditional GWAS and MTAG. Analyses of serum metabolomics data presented in the fourth chapter revealed widespread metabolic rewiring, reflecting glycine and serine commensurate with genetic disruptions and identified lipids dysregulation as a new risk factor for MacTel. In the fifth chapter, we confirm that genetically induced depletion of serine in MacTel patients likely results in accumulation of deoxy-sphingolipids. These appear to be extremely toxic for retinal and neuronal tissues and were associated with disease

progression. Lastly, the sixth chapter summarises and discusses the main findings of these four studies while placing their discoveries on the bigger spectrum of macular and eye disorders as well as suggesting future studies. The results presented in this thesis significantly advance the understanding of MacTel aetiology and are supporting sensitive diagnostic re-evaluation and targeted laboratory experimentation as well as inclusion criteria for upcoming clinical trials.

Declaration

This is to certify that

i. The thesis comprises only my original work towards the degree of Doctor of Philosophy except where indicated in the Preface,

ii. Due acknowledgement has been made in the text to all other material used,

iii. The thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed,

Roberto Bonelli

Preface

Several chapters of work in this thesis were carried out in collaboration with others.

Part of the results from Chapter 2 have been published in the following:

Scerri, T. S., Quaglieri, A., Cai, C., Zernant, J., Matsunami, N., Baird, L., Scheppke L., **Bonelli R**. et al., (2017). Genome-wide analyses identify common variants associated with macular telangiectasia type 2. Nature Genetics. https://doi.org/10.1038/ng.3799.

The specific contributions of Roberto Bonelli to this specific publication were the heritability analysis, eQTL analysis, disease prediction analysis using SNP data and the semi-targeted metabolomics analysis. Roberto Bonelli additionally contributed to the interpretation of the results and the drafting of the original manuscript.

The primary collaborators involved in this work were Professor Melanie Bahlo who directed the study, conceived ideas and provided guidance. Dr Thomas Scerri who performed most of the GWAS analysis and wrote the manuscript. Ms Anna Quaglieri who performed several statistical analysis. I performed all statistical analysis for the heritability study, metabolomics study, eQTL study and penalised regression prediction.

The results from Chapter 3 are confidential, to be considered in embargo, and are submitted to Genetics in Medicine as the following:

"**Bonelli R.** et al. Genetic Disruption of Serine Biosynthesis is a Key Driver of Macular Telangiectasia Type 2 Aetiology and Progression".

Chapter 3 contains this manuscript as reviewed by all co-authors in the same format as intended for submission to peer-reviewed journals.

Roberto Bonelli contributed to the design of this study, performed all the statistical analysis, contributed to the interpretation of the results, and wrote the manuscript.

The primary collaborators involved in this work were Professor Melanie Bahlo who conceived the study and provided guidance. Dr Brendan Ansell who provided guidance, helped interpret the results, and helped write the manuscript. Dr Luca Lotta who provided the data needed for the Mendelian Randomization study and provided guidance on its usage. Dr Claudia Langerbergh who provided guidance and the Mendelian randomization data. Dr Ferenc Sallo who provided the retinal phenotypic data guidance, its usage and helped to interpret the results. Dr Traci Clemons who provided the retinal phenotypic data and provided guidance on its usage. I conducted all the analysis and wrote the manuscript, constituting 80% of the work.

The results from Chapter 4 are confidential, to be considered in embargo, and have been submitted and already peer-reviewed by Scientific Reports as:

"**Bonelli R.** et al, Systemic lipid dysregulation is a novel risk factor for macular neurodegenerative disease".

Chapter 4 contains this manuscript as reviewed by all co-authors in the same format as intended for submission to the journal Scientific Reports.

Roberto Bonelli contributed performed all the statistical analysis, contributed to the interpretation of the results, and wrote the first draft of the manuscript.

The primary collaborators involved in this work were Prof. Marcus Fruttiger who conceived the study provided the metabolomics data guide on its usage, interpreted the results, and partially wrote the manuscript. Prof. Melanie Bahlo who conceived ideas and provided guidance on the statistical analysis. Dr Brendan Ansell who helped to interpret the results and helped write the manuscript. Dr Sasha Wood who produced the metabolomics data, helped interpret the results and helped write the manuscript. Dr Catherine Egan who helped to conceive the study, and helped to collect the sample for metabolomics processing. I conducted all the analysis and wrote the initial version of the manuscript, constituting 70% of the work.

Some of the results from Chapter 5 have been recently published in the New England Journal of Medicine as the following:

Marin L. Gantner, Kevin Eade, Martina Wallace, Michal K. Handzlik, Regis Fallon, Jennifer Trombley, **Roberto Bonelli**, et al. Serine and Lipid Metabolism in Macular Disease and Peripheral Neuropathy. The New England Journal of Medicine. https://doi.org/10.1056/NEJMoa1815111.

Roberto Bonelli performed some of the statistical analysis, in particular, he analysed the human and mouse plasma metabolic levels, as well as providing statistical consultation for the analysis performed on cell culture and mice experiments. Roberto Bonelli additionally contributed to the interpretation of the results the manuscript editing.

The primary collaborators involved in this work were Dr Marin Gantner who conceived the study, performed the mouse experiments, interpreted the results, and wrote the manuscript, Dr Martina Wallace who conceived the study, performed the metabolomics measurements, interpreted the results, and wrote the manuscript, Dr Kevin Eade who conceived the study, performed the cell culture measurements, interpreted the results, and wrote the manuscript, Dr Christian Metallo who conceived the study, provided guidance and interpreted the results, Dr Martin Friedlander who conceived the study, provided guidance, and interpreted the results, Prof. Melanie Bahlo who provided guidance on the statistical analysis. I performed most of the statistical analysis of the manuscript, checked the statistical methodologies involved and helped write the manuscript.

Throughout this thesis an unpublished manuscript will be cited several times. This is indicated by the citation "Lotta et al". This study will be submitted within the year 2019 at Nature Genetics.

This thesis comes with supplementary materials specific to Chapter 3 and Chapter 4. These supplementary materials are provided as digital documents given their large size.

Additional work was performed as part of this PhD that is not included in this thesis. This includes the following MacTel observational studies;

1. Heeren, T. F. C., Tzaridis, S., **Bonelli, R**., Pfau, M., Fruttiger, M., Okada, M., … Holz, F. G. (2019). Dark-Adapted Two-Color Fundus-Controlled Perimetry in Macular Telangiectasia Type 2. Investigative Ophthalmology & Visual Science, 60(5), 1760–1767.

2. Leung, I., Sallo, F. B., **Bonelli, R**., Clemons, T. E., Pauleikhoff, D., Chew, E. Y., … MacTel Study Group. (2017). Characteristics of pigmented lesions in type 2 idiopathic macular telangiectasia. Retina. https://doi.org/10.1097/IAE.0000000000001842

3. Müller, S., Heeren, T. F. C., **Bonelli, R**., Fruttiger, M., Charbel Issa, P., Egan, C. A., & Holz, F. G. (2018). Contrast sensitivity and visual acuity under low light conditions in macular telangiectasia type 2. The British Journal of Ophthalmology. https://doi.org/10.1136/bjophthalmol-2017-311785

4. Okada, M., Heeren, T. F. C., Egan, C. A., Rocco, V., **Bonelli, R**., & Fruttiger, M. (2017). Effect of dark adaptation and bleaching on blue light reflectance imaging in macular telangiectasia type 2. Retina. https://doi.org/10.1097/IAE.0000000000001754

5. Pauleikhoff, D., **Bonelli, R**., Dubis, A. M., Gunnemann, F., Rothaus, K., Charbel Issa, P., … Zhuk, S. A. (2019). Progression characteristics of

ellipsoid zone loss in macular telangiectasia type 2. Acta Ophthalmologica, 38, S20.

Roberto Bonelli performed less than 50% of the work in each of these publications.

Acknowledgments

First and foremost, I would like to thank my main supervisor Melanie Bahlo who has guided me with extreme patience, kindness, and understanding throughout this journey. Melanie, thank you for the countless opportunities you have given me. You are and have been an endless source of inspiration, both as a scientist and as a leader. I could not have asked for a better supervisor.

Next, I would like to thank my mentor, colleague, and above all, friend, Saskia. Saskia you have been my mentor and second supervisor the second I set foot into WEHI. Thank you for all you have done for me, for you many lessons, for your patience, for the endless chats and most of all for your incredible friendship.

I would like to thank my other colleague, mentor and supervisor Brendan. Thanks, Brendan for being so patient and so excited about everything we have done together. Thanks for all the advice you have given me, working with you has been an absolute pleasure and something I surely look forward doing again in the future.

I would like to thank who has become more like a sister than a colleague/friend, Anna. Anna, I've known you for 10 years now and we have shared most of our academic experience together. Working, living, chatting, and sharing a third of my existence with you has bettered me in ways I cannot even begin to explain. Thank you for literally everything we have shared together.

I'd like to thank my PhD committee members Terry, Matt, Oliver, and Traci. Thanks for all your support, advice and kind words throughout my PhD.

I would like to thank all the past and present members of the Bahlo Lab whose collaboration, friendship and support has been pivotal for this thesis. In particular

I'd like to Brendan, Vicky, Mark and Haloom for they precious help and feedback on the initial draft of this thesis.

I would like to thank all members of the MacTel consortium. In particular, I'd like to thank Rando, Marcus, Sasha, Mari, Martina, Kevin, Christian, Martin, Cathy, Tjebo, Ferenc, Irene, Tunde, and Marti Lynn. Thank you for all your feedback, knowledge, guidance and opportunities you have given me throughout this journey. It has been an extreme pleasure and privilege to work with you all.

I'd like to thank our Cambridge University collaborators, Luca and Claudia and their team for their support and guidance.

I thank the MacTel Consortium, the Lowy family, the Melbourne International research scholarship, the University of Melbourne, The Walter and Eliza Hall Institute, Edith Moffat, and the John and Patricia Farrant foundation, whose funding and generosity has enabled me to fully devote my time to this research, to travel overseas and explore research institutes.

I'd like to thank WEHI directors Doug and Sam for their support to both my academic and social work. Thanks to your leadership I could not have found a better, more inclusive, and more inspiring place to do a PhD.

I would like to thank all WEHI services teams without whom my work would not have been even remotely possible. In particular, I'd like to thank Louise J, Louise N, Chris, Lucy, Arunee, Cal, Peter, Rachel, Ellen, Jaci, Simon, Rosie and many others whose names would fill the rest of this thesis for their incredible help and support during the last years.

A special thanks goes to Medu and Figen, who always had a kind word for me and prevented me from starvation for half of my PhD.

I'd like to thank all my fellow WEHI student association committee members for the incredible work done together.

To all my amazing WEHI friends, thanks for everything you have done, all the chats, the laughs, the parties, and the lunches. I had an absolute blast during my PhD and its mostly thanks to you all.

I would like to thank the QueersInScience committee, the WE-Pride group, and in particular, to Sarah S. You guys have made possible a dream I could have not thought possible and you are all making the world a better place.

I would also like to thank all current and previous housemates Steena, Gina, Steven, Riccardo, Gerry, Lauren, John, Jan, Andres, Lucille, Utkarsh, Prasad, Kristen and Annalise. You have been my family for the past 5 years. Living with you and be your friend has been incredible, it has changed and improved me, and is most likely the main reason why I still retain some sort of mental sanity. I will never forget any of you (apart from your names).

To all my friends in Australia, thanks for everything you have done and for everything you have shared with me.

Thanks to my partner Mark, whose patience, kindness, and love has inspired me, supported me, and prevented me from going insane during the last months.

Last, but most certainly not least, to my family Mauro, Enza and Simone and to my Italian friends Federico, Christian, Alberto, Carolina, Alice, Elena, and Nello. You guys have made Roberto. You have allowed me to become the person I am today. Thanks for all your love, support and presence during all these years even though we were separated by an entire planet.

To all of you, again, Thank You!

Contents

List of Figures

* The following list of figures does not contain indexing of the images used in the manuscripts presented in chapter 3.2 and 4.2. Indexing of figures in such chapters are to be considered separate and refers only to the presented manuscript.

List of Tables

Abbreviations

Ala	Alanine
AMD	Age-related Macular Degeneration
AUC	Area Under the receiving operating Curve
Bp	Base Pair
Cer	Ceramide
Chr	Chromosome
CNF	Ciliary Neurotrophic Factor
Cys	Cystine
doxSL	Deoxy-Sphingolipid
DR	Diabetic Retinopathy
DNA	Deoxyribonucleic acid
eQTL	Expression Quantitative Trait Loci
EZ	Ellipsoid Zone
FA	Fluorescein Angiography
FC	Fold-Change
GLM	Generalised Linear Model
Gly	Glycine
GP Met	Glycerophospholipid Metabolism
GTEx	Genotype-Tissue Expression (consortium)
GWAS	Genome-Wide Association Study
KEGG	Kyoto Encyclopedia of Genes and Genomes

Kbp	Kilo Base Pair
HSAN1	Hereditary Sensory Neuropathy Type 1
ICD-10	10th revision of the International Statistical Classification of Diseases and Related Health Problems
IS	Inner Segment
LD	Linkage Disequilibrium
LM	Linear Model
LMM	Linear Mixed Model
M	Million
MacTel	Macular Telangiectasia Type 2
MAF	Minor Allele Frequency
Met	Metabolism
mQTL	Metabolic Quantitative Trait Loci
MR	Mendelian Randomization
MTAG	Multi-Trait Analysis of GWAS
NHOR	Natural History Observational Registry
NHOS	Natural History Observational Study
OCT	Optical Coherence Tomography
OS	Outer Segment
QC	Quality Control
QTL	Quantitative Trait Loci
PC	Principal Component
RB	Roberto Bonelli
RPE	Retinal Pigment Epithelium
RNA	Ribonucleic Acid

RCT	Randomized Clinical Trial
SA	Sphinganine
Ser	Serine
SNP	Single Nucleotide Polymorphism
Sph	Sphingomyelin
SPT	Serine C-palmitoyltransferase
T2D	Type 2 Diabetes
Trp	Tryptophan
Tyr	Tyrosine
UPLC-MS	Ultrahigh-Performance Liquid Chromatography-tandem Mass Spectroscopy
Var	Variance
VGEF	Vascular Endothelial Growth Factor

1 Introduction

The following chapter will explore the main concepts required to understand the framework of this thesis. We will describe the disease that is the target of this work: Macular Telangiectasia Type 2, also known as MacTel, and what is known about it. We will then explore how research on a disease such as MacTel can be conducted in the current era of 'omics data and describe the types of datasets used in this thesis. Lastly, we will present different methodological frameworks that will be used throughout the rest of this thesis.

1.1 Macular Telangiectasia Type 2

1.1.1 Description

Macular telangiectasia type 2, typically abbreviated as MacTel, is an eye disease. This disease affects the macula, the central area of the retina which has the highest density of photoreceptors in the human eye. Initially described by Donald Gass in 1977 (1), it is a bilateral disease of mainly unknown causes (2). Depending on the stage of the disease it reduces visual acuity (3), reading ability (4), vision-related quality of life (5, 6) and distorts vision (2).

1.1.2 Epidemiology

The disease prevalence has been estimated to be between 0.004% (7) and 0.1% (8). Given the rarity of the disease and the subtlety of its clinical signs, the disease

was initially largely under- and misdiagnosed (2) and to this day remains challenging to identify. In 2005 a privately funded project, called the "MacTel Project", started a longitudinal cohort study of MacTel patients, also known as the "MacTel Natural History Study". The aim of the project was to better understand the development of the disease and explore its natural history (9). Five years later, Clemons et al (10) described the cohort of the MacTel Natural History Observational Study (NHOS). Gender and ethnicity were not found to be risk factors for this disease. However, age was identified as playing an important role in the disease manifestation. In the MacTel cohort study, the mean age at diagnosis was 57 years identifying MacTel as a late-onset disease. In a subsequent publication, Clemons et al. (11) explored additional risk factors presented by the subjects enrolled in the NHOS. Interestingly, there was a significant difference between MacTel cases and controls in the prevalence of heart disease and hypertension. These two conditions are associated with increased BMI and weight, which were also identified as risk factors in the MacTel cohort. Cancer prevalence was also suggestively significantly increased in the NHOR study of MacTel patients, and 50% of the individuals in the cohort were either current or former smokers. However, which of these correlations are causal for MacTel and which are due to ascertainment bias or confounding was unknown.

The most striking difference between the MacTel patients and the healthy cohort in this study was the high prevalence of diabetes mellitus (T2D) (38.2% compared to 10% of controls) among MacTel patients. This association was concordant with a previous study conducted in India (12) which reported a prevalence of (59% T2D

among 104 MacTel patients). Interestingly, in both studies, there was almost no sign of subjects suffering from diabetic retinopathy. This is a frequent comorbidity in T2D and some characteristics of this disease resemble MacTel disease. The absence of diabetic retinopathy in conjunction with the higher prevalence of diabetes mellitus in the MacTel cohort has suggested that the genetic component behind MacTel may be protective against diabetic retinopathy (11).

1.1.3 Clinical signs

Clinical signs of MacTel disease are observed in the *retina*. In humans, the retina is situated on the back of the eye and is the most light-sensitive tissue in the eye. An image of a healthy retina is shown in **Figure 1** (13):

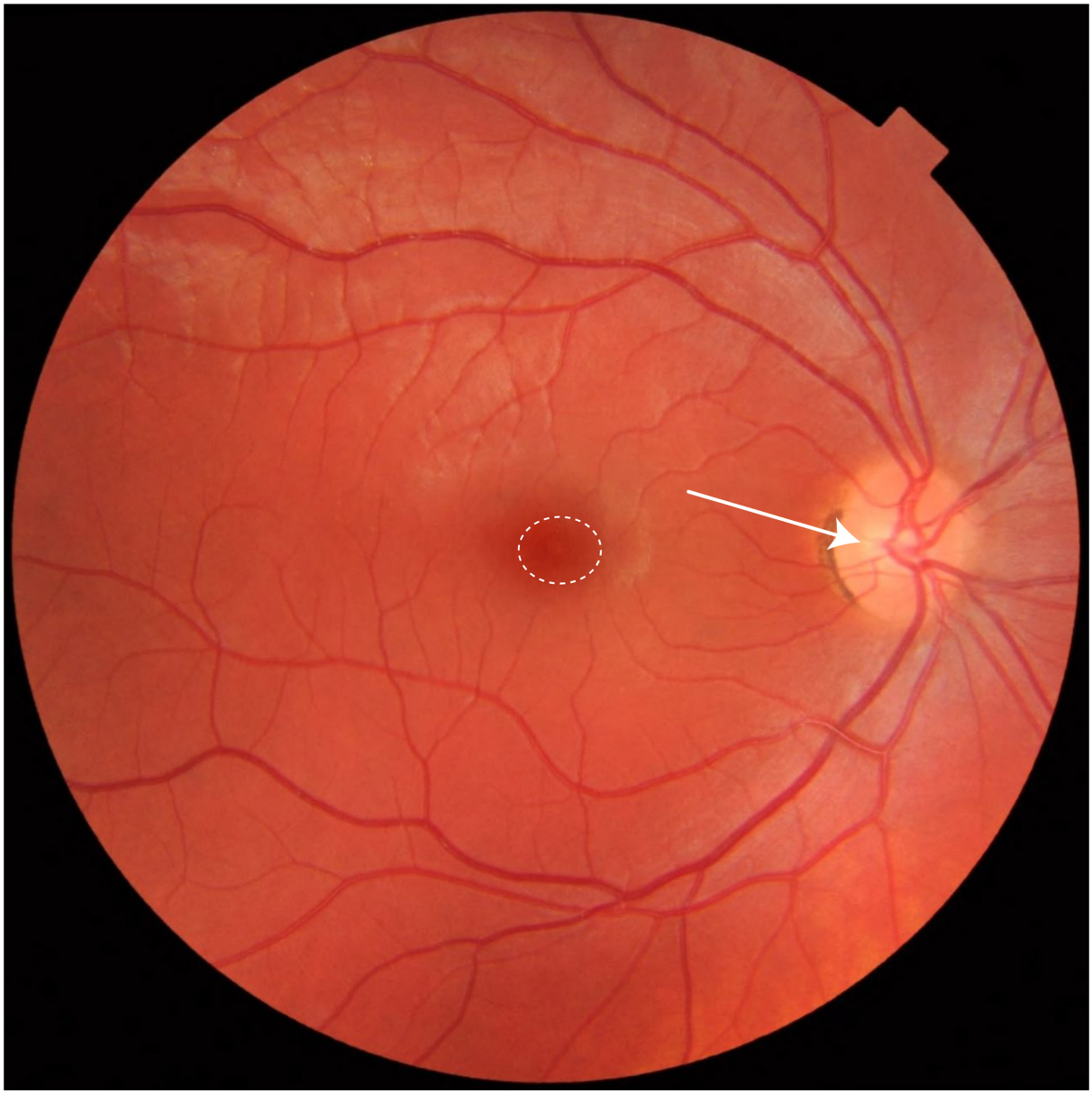


Figure 1: A normal retinal colour fundus image. The dashed circle indicates the area of the retina known as the macula, while the white arrow points to the optical nerve. This mage was received through internal collaboration.

The site of lesion for this disease is a region at the centre of the retina called the *macula*. In **Figure 1**, the macula, outlined here with a dashed circle, is generally identifiable as the darker round area in the middle of the retina. The bright yellow

circle on the right side of the image is the *optical nerve,* which is not affected in MacTel disease.

In human vision, the macula is extremely important since it contains most of the *photoreceptors* present in the eye. Photoreceptors are particular neuronal cells, responsible for vision, which react to the light entering the eye and transmit that information to the brain. The mouse, a key experimental animal for biomedical research, lacks a macula, as do rats. The very centre of the macula is called the *fovea.* The fovea contains the photoreceptors responsible for the central vision, the area of the eye crucial for visual acuity, or fine-scale vision which helps humans read, recognise faces etc. There are two sides of the retina, the *nasal* side and *temporal* side. The nasal side is the area where the optic nerve is visible (right side in **Figure 1**), and to which side the nose is in the human face (hence nasal), while the temporal side is the opposite side (left side in **Figure 1**).

Clinical signs of MacTel initially appear temporal to the fovea (2) and are usually symmetric, being present in both eyes (10), making MacTel a bilateral and symmetric disease. Patients affected by MacTel present with a variety of clinical signs in the macula that may vary with the disease course.

Different ophthalmological diagnostic techniques are used to identify MacTel clinical signs, such techniques will be explored here since data from these instruments are used later in the thesis.

1.1.3.1 Colour fundus photography

Various macular alterations can be found using *colour fundus photography* which essentially takes a coloured picture of the retina.

One of the earliest signs in MacTel is the *loss of retinal transparency* (2) (**Figure 2**.

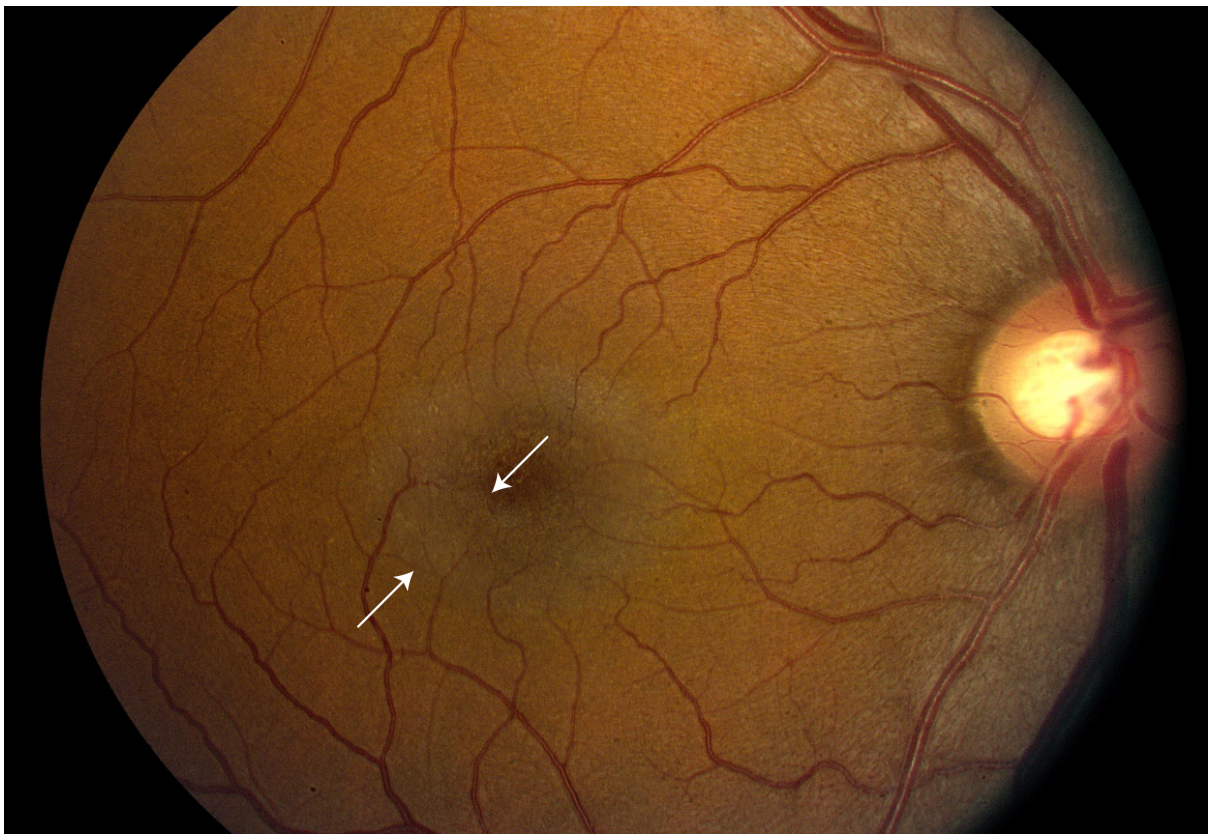


Figure 2: Loss of retinal transparency as observed in colour fundus imaging on retina affected by MacTel. Image received through internal collaboration.

The loss of retinal transparency can be seen as the grey area indicated by the arrows. This grey area starts temporal to the fovea and progresses to cover the entire circle around the macula.

Another common sign, sometimes visible in colour fundus photography, is the presence of blunted and dilated vessels, as well as right angle veins (**Figure 3**).

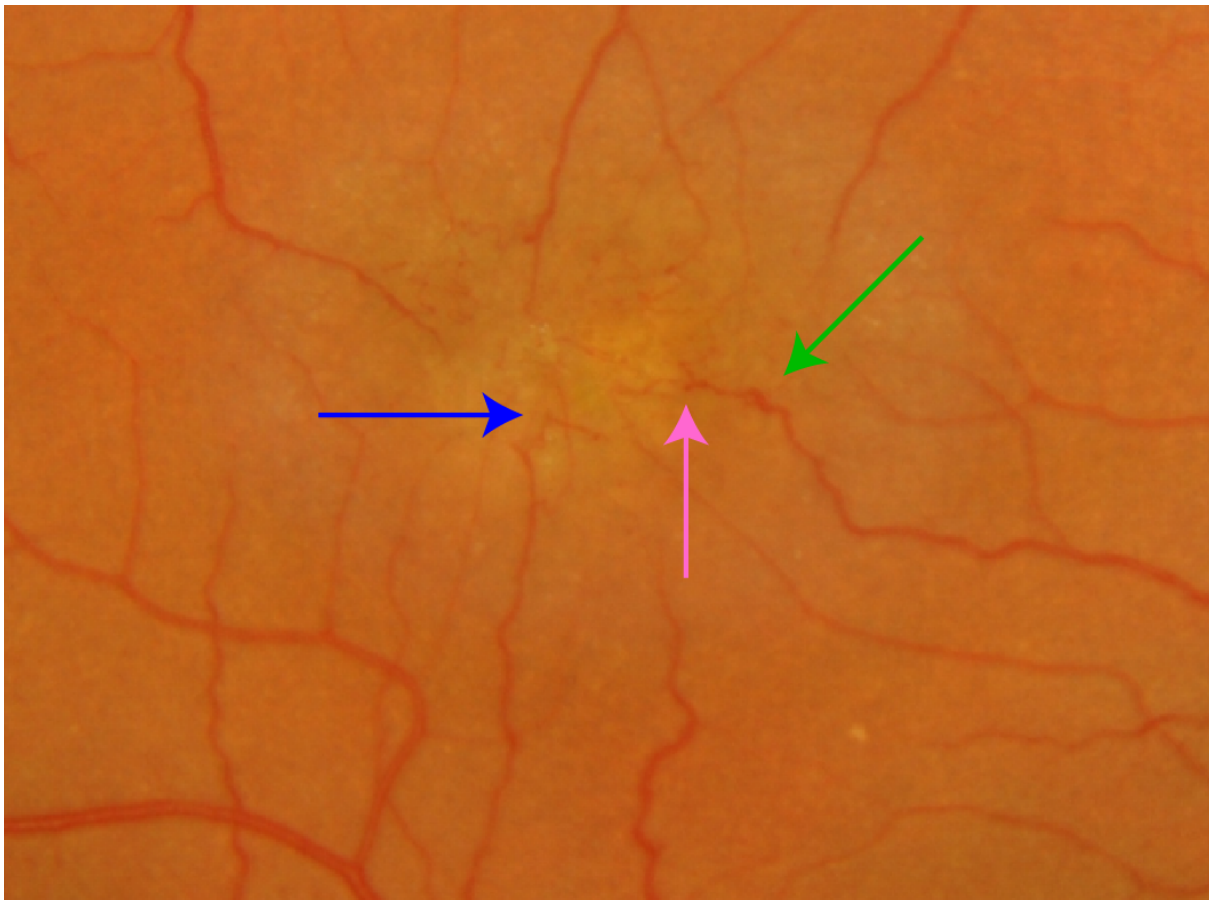


Figure 3: Vasculature signs typically observable from colour fundus imaging in retinas affected by MacTel. Image received through internal collaboration.

Vessels in retina should progressively narrow the more they converge to the centre. In **Figure 3** the green arrow indicates a dilated vessel which can be seen to be enlarged. The same vessel can also be seen with the pink arrow to be blunt at its end, near the macula. As already mentioned, healthy vessels should narrow progressively until disappearing from sight while converging towards the fovea. However, it can be noted that the blunted vessel suddenly disappears after a small

enlarged area. This is caused by the vessel "diving" down towards the *outer* retinal layers instead of remaining in the surface. The term "outer" is used in this context since the retina observed through these images is facing the *inner* side of the eye. Lastly, the blue arrow indicates right-angle veins. These are simply veins that instead of progressively converge towards the centre suddenly turn creating an "angle". This is also thought to be caused by the vessels suddenly diving into the retina. The process that leads to these abnormal blood vessels is called *telangiectasia* and blood vessels that show it are referred to as being *telangiectatic*.

Another clinical sign of MacTel is defined by the presence of retinal pigment plaques as in **Figure 4**.

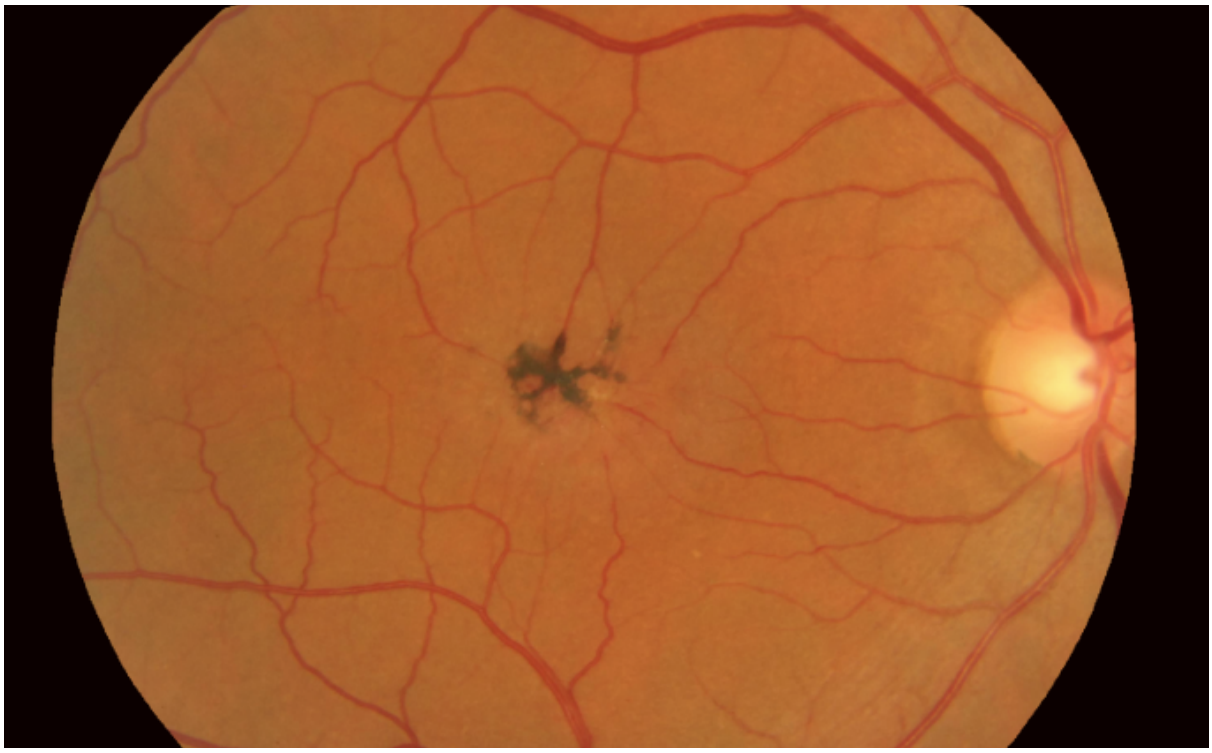


Figure 4: Pigment epithelium migration plaques as shown by fundus colour imaging in retina affected by MacTel. Image received through internal collaboration.

In **Figure 4** a large pigment plaque can be seen as a type of black "stain" in the macula. This plaque is the result of the proliferation of pigmented cells rising through the surface from the *Retinal Pigment Epithelium* (RPE) layer. The current hypothesis behind this phenomenon is related to blunted and right-angled veins (14). In fact, a vessel that has dived inside the retinal outer layers will eventually meet the RPE layer. Once the vessels are in contact with the RPE, the pigment cells will use them to rise towards the inner retina.

Another sign of the disease visible from colour fundus photography is the presence of crystals around the macula as shown in **Figure 5**.

Figure 5: Crystalline deposits as shown in colour fundus imaging of retina affected by MacTel. Image received through internal collaboration.

However, the presence of crystals is an uncommon sign for this disease and it is thought to be related to a metabolic disturbance in the eye (2).

1.1.3.2 Fluorescein Angiography

A specific eye examination called *Fluorescein Angiography* (FA), is the current gold-standard imaging method to confirm the diagnosis of MacTel. This examination consists of an intravenous injection of fluorescein dye. A fundus camera equipped with specific lenses that react to the colour emitted by the dye is then placed in front of the eye.. The camera will capture images of the retina immediately after and for another 10 minutes following the injection, revealing all blood vessels in the retina and any potential leakage.

Fluorescein angiography of a retina affected by MacTel disease is displayed in **Figure 6**.

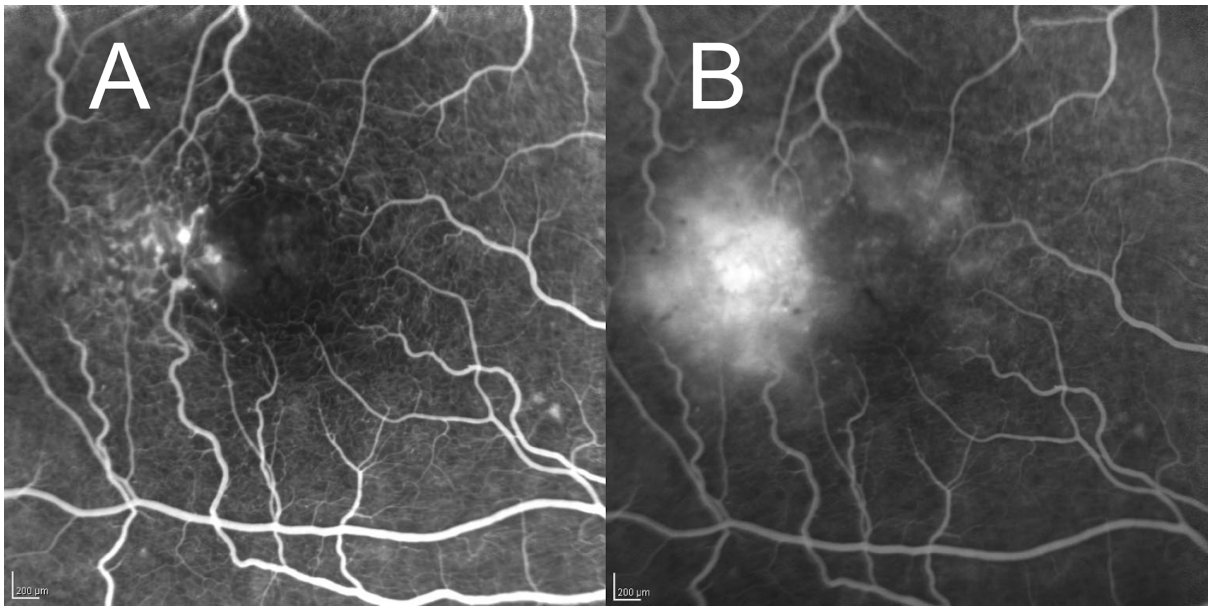


Figure 6: Fluorescein angiography of a retina affected by MacTel, shortly after fluorescein injection (A), and long after fluorescein injection (B) (~10-minute mark). Image received through internal collaboration.

Figure 6 displays the results of two fluorescein angiography photographs. One immediately after the injection of the dye (A), and the other some minutes after injection (B). In **Figure 6** (A) we can observe telangiectatic vessels temporal to the fovea (left of the fovea). These vessels are dilated and some of them blunted, or even right-angled. From this figure, it is already possible to notice how these vessels are prominent to leakage of the dye. Additionally confirming the leakage is **Figure 6** (B), where the dye can be seen to have spread around the macular tissue.

1.1.3.3 Optical Coherence Tomography (OCT)

Optical coherence tomography is a widely used technique that allows a cross-sectional visualisation of the retinal layers. A healthy retina's OCT is shown in **Figure 7** adapted from (15).

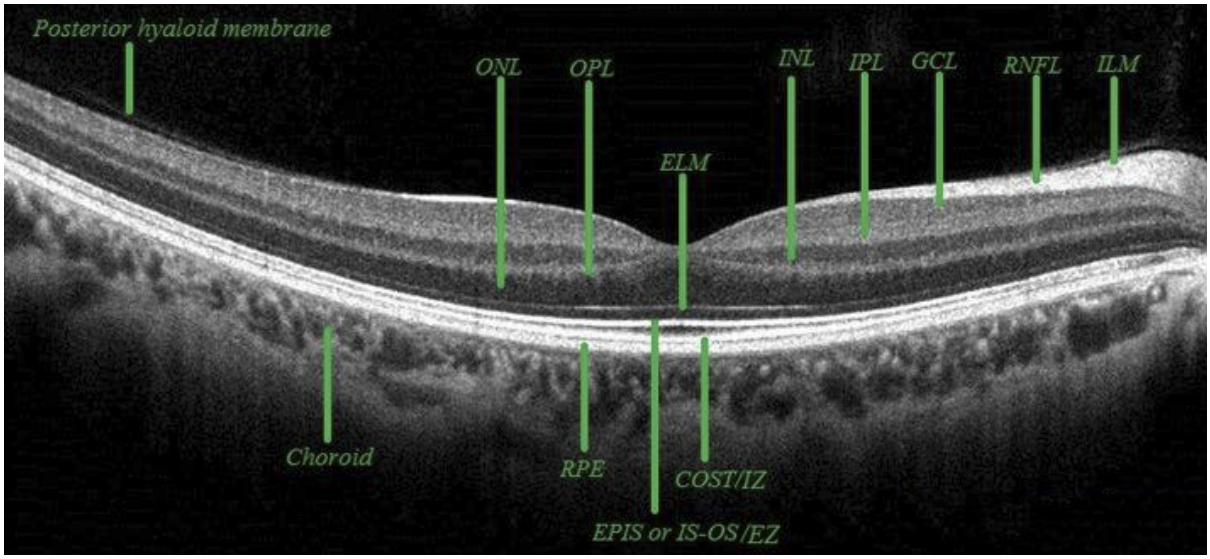


Figure 7: Healthy retina anatomy as shown by OCT imaging. Layers of the retina are indicated by a green vertical line. The fovea is visible as a depression in the middle in the macula (behind ELM green arrow). ONL: Outer Nuclear Layer. OPL:

Outer Plexiform Layer. INL: Inner Nuclear Layer. IPL: Inner Plexiform Layer. GCL: Ganglion Cell Layer. RMFL: Retinal nerve fibre layer., ILM: Inner Limiting Membrane. ELM: External Limiting Membrane. IS-OS/EZ: Inner Segment - Outer Segment layer / Ellipsoid Zone. CIST: Cones Outer Segment Tips layer. RPE: Retinal Pigment Epithelium layer.

A comparison between a healthy and MacTel affected retina under OCT scan is presented in **Figure 8**.

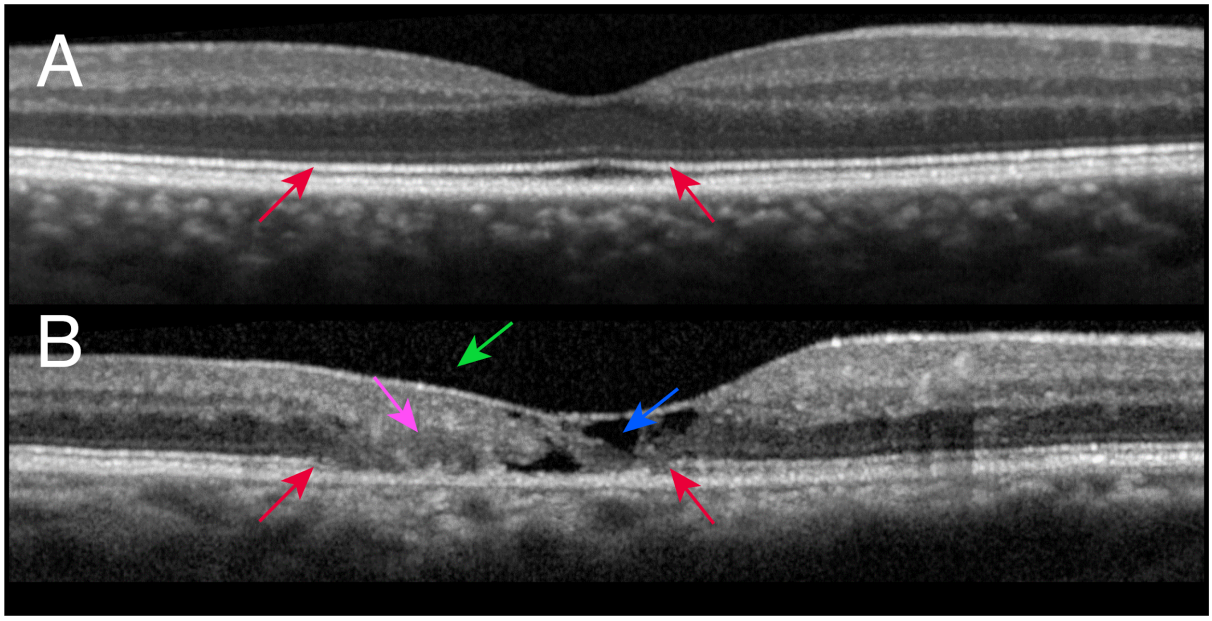


Figure 8: Comparison of OCT MacTel clinical sign between healthy retina (A) and affected retina (B). Image received through internal collaboration.

One of the layers affected by MacTel is the IS/OS junction layer which divides the inner photoreceptors layer from the outer photoreceptors layer. Comparing the red arrows in **Figure 8** A and B shows how in the MacTel affected retina this layer tends to reduce until disappearing. For this reason, this clinical sign is referred to as the *Ellipsoid Zone Loss* (EZ loss). We explored the behaviour of this clinical sign

and its progression in a recent publication where we noticed that this was exponential in its decay (16). Other early signs of the disease are the flattening of the temporal foveal slope (green arrow), leakage due to telangiectatic vessels resulting in retinal cystoids (blue arrow), and a general degradation and collapse of the retinal layers (pink arrow).

Although not presented in the previous figure, other clinical signs often visible in OCT scans are the degradation of the RPE layer, subretinal vessel and macular/lamellar holes.

An important phenotype measurable from OCT scans is the retinal thickness. The thickness is defined as the distance at each time point between the inner limiting membrane (the very top of the OCT scan) and the outer layer the varies by the technology used (17). In eyes affected by MacTel disease, a reduction in the macular thickness compared to healthy eyes is observed, due to the degradation of different layers in the retina.

1.1.3.4 Fundus autofluorescence, blue light reflectance, and dual wavelength light

Many other ophthalmological imaging techniques have been used to explore the clinical signs of MacTel disease. Three other common techniques are shown in **Figure 9**.

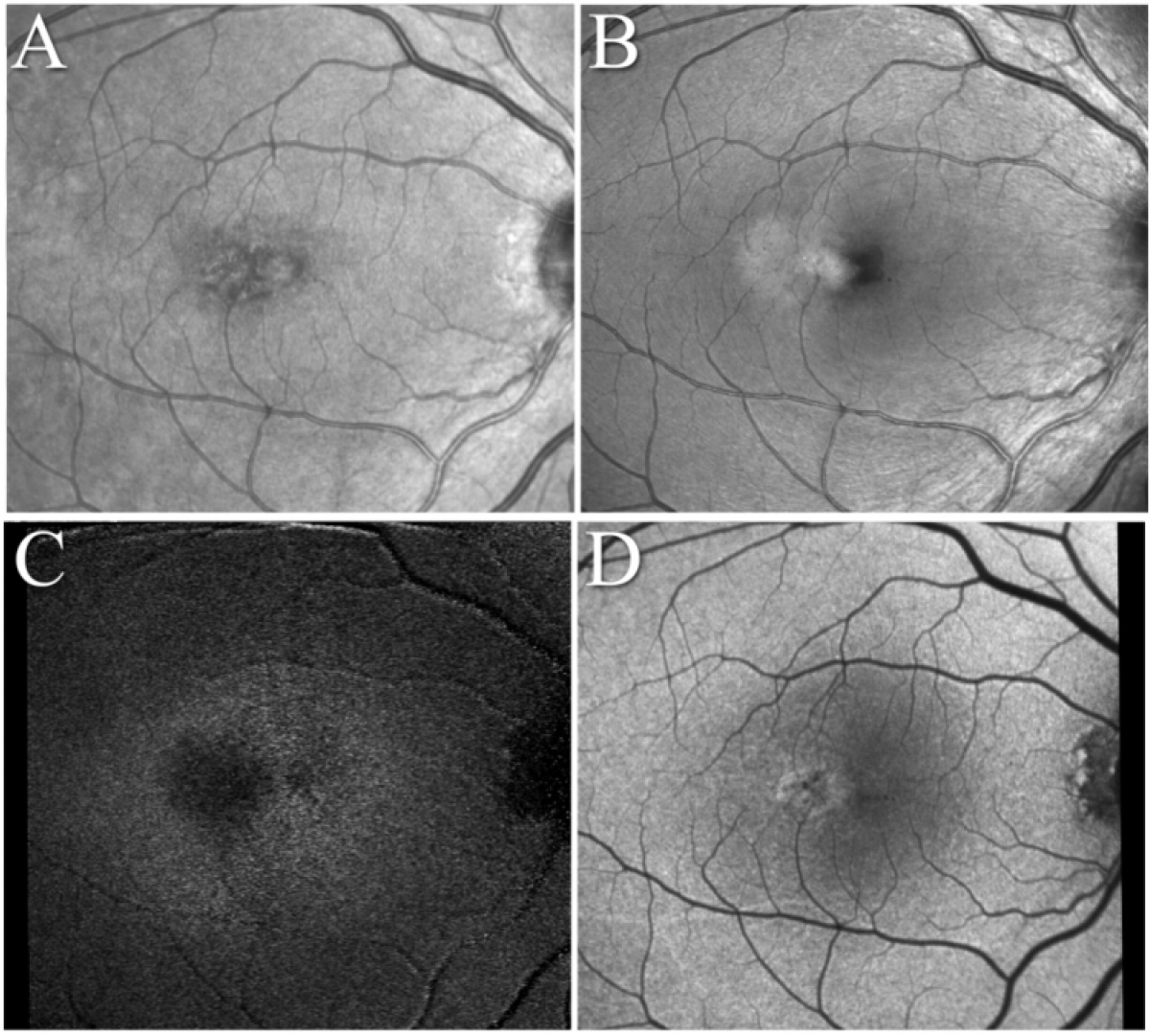


Figure 9: Results of blue light reflectance (A-B), dual-wavelength autofluorescence (C), and fundus autofluorescence (D). Image received through internal collaboration.

Figure 9 A and B are the result of blue light reflectance imaging. In this test, the loss of retinal transparency is clearly highlighted. **Figure 9** C presents a dual-wavelength autofluorescence image which is able to identify the loss of macular pigment, identifiable in this figure as the darker area temporal to the fovea. Not much is known about this clinical sign although it often overlaps with loss of retinal transparency. Lastly, **Figure 9** D shows the results of fundus

autofluorescence imaging. This technique is often used to assess the health status of blood vessels and the RPE which, when damaged, appears as a brighter area, as in this figure.

1.1.4 Staging

MacTel progresses slowly over time, with worsening of some of the clinical signs, or with the appearance of new clinical indicators (2).

The progression of the disease was initially divided into five different stages by Gass and Blodi in 1993 (3), depending on the clinical signs present in the patients' retinas. This staging system was created when the understanding of this disease was very limited and only colour fundus photography and fluorescein angiography techniques were available. The five stages were defined as follows:

1. Stage 1: hyper-fluorescence in fluorescein angiography usually temporal to the fovea indicating small leakages of the vessels in the retina.
2. Stage 2: loss of retinal transparency highlighted by a greying of the parafoveal region on fundus colour imaging. Mild telangiectatic vessels might be seen in this stage.
3. Stage 3: strong telangiectasis defined by dilated, blunted and right-angled vessels. These vessels can belong to both the venular or the arterial system.
4. Stage 4: retinal plaques caused by the proliferation to the surface of RPE.
5. Stage 5: neovascularization.

Additional to the age of this classification system, stages 1-3 of this MacTel classification system have been characterised by very low inter-observer reliability

(2), however, this system is still used to this day to define disease progression and will be used throughout this thesis.

Interestingly, it has been noted that neo-vascularization can occur in most of the stages, with different intensities (2). For this reason, neo-vascularization should not be considered the natural endpoint of the disease. To solve this problem MacTel has been divided into two subcategories: proliferative and non-proliferative. The proliferative category is characterised by neovascularization, which is absent in the non-proliferative category.

The natural endpoint of the disease is the inevitable atrophy of the photoreceptors layer (2). This atrophy is caused by the death of the photoreceptor cells contained in the retina which is the direct consequence of all the retinal abnormalities involved in the disease. Since photoreceptors are the cells which react to the light and send a signal to the brain in order to create the vision, absence of photoreceptors directly translate into vision loss.

1.1.5 Vision Loss

It has been observed that patients initially lose their peripheral vision and first report to their ophthalmologists with signs of scotoma and metamorphopsia. The former indicates the presence of a “blind spot” while the latter signifies a distortion of the image perceived by the eye. Metamorphopsia has been observed to appear before the presence of scotoma (2). Scotomas are measured by an ophthalmological examination called microperimetry which tests whether light projected into

different parts of the retina is perceived by the subjects. An example of microperimetry performed on three patients' right eyes along different time points is presented in **Figure 10**.

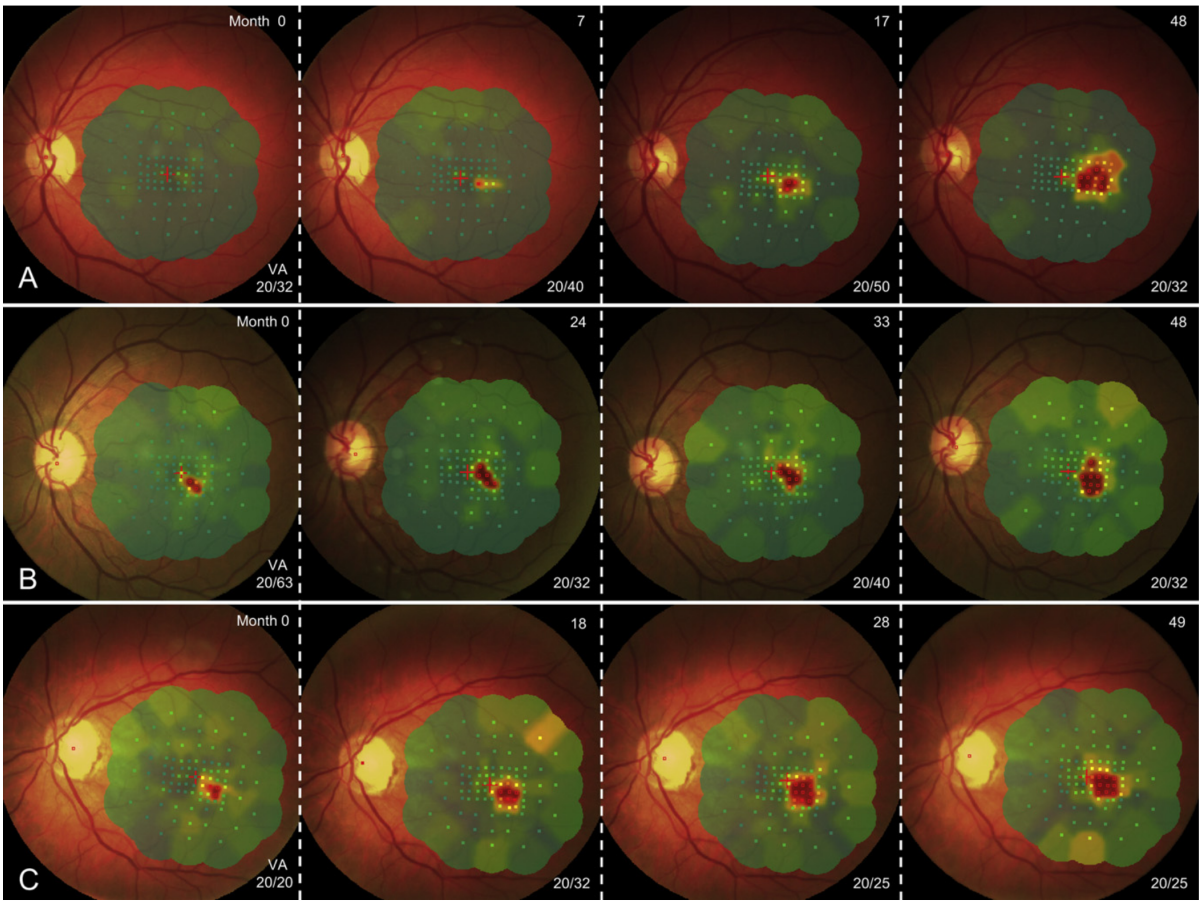


Figure 10: Example of microperimetry output on subjects affected by MacTel. The three patients are displayed separately in each row. The columns represent different time points, with scotomas (red areas) increasing in size (as measured by microperimetry) over time. Image received through internal collaboration.

In these three patients, scotomas start on the temporal area of the fovea (central red cross) to then progress around it. Reading ability and especially reading speed is often reduced in MacTel patients. This is the direct consequence of the

paracentral scotoma that, by impairing the peripheral vision, makes it difficult to recognise subsequent letters or to follow the same written line (4). Central vision and hence visual acuity are only affected in very late stages (from stage 3) of MacTel. However, in a recent study, we noted that central vision of contrast sensitivity testing using the Pelli-Robson chart shown in **Figure 11** can detect very early stages of MacTel disease in mesopic conditions (a room with the low light condition) (18).

Figure 11: Sample of the Pelli-Robson chart used to test contrast sensitivity.

1.1.6 Therapeutic approaches

Various therapeutic treatments have been tested in the past to treat this disease. Given that neo-vascularization can appear at any stage of the disease, therapeutic

approaches have been tried on both non-proliferative and proliferative disease (19). For the non-proliferative MacTel, argon laser photocoagulation, photodynamic therapy, and intravitreal injection of steroids have been attempted as therapies. None of these provided evidence of any significant improvement on the disease. Furthermore, some of these methods have been thought to accelerate progression of MacTel.

Non-proliferative MacTel disease is characterised by different vessels abnormalities including, vessel leakage, and blunted and ectatic vessels. For this reason, it was assumed that VEGF inhibitors could be used as a potential therapeutic method. Early observations showed evidence of reduction of vessel abnormalities of the patients treated by regular intravenous injections of anti-VEGF. However, long-term observations demonstrated poor performance (if not an even worse effect) on disease progression and visual impairment (20). For this reason, VEGF inhibitors have been excluded for the treatment of non-proliferative MacTel disease.

For proliferative MacTel disease therapeutic approaches such as focal laser photocoagulation, photodynamic therapy, transpupillary thermotherapy, posterior juxtascleral administration of steroids, and intravitreal injection of VEGF-inhibitors have been tried. Most treatments were found to have an effect on the early stages of the vascular membrane and proliferation. However, only anti-VEGF showed an improvement on all vascular symptoms.

Recently, a phase 1 (21) and a phase 2 (22) clinical trial have been conducted to test the effect of ciliary neurotrophic factor (CNTF), delivered via encapsulated retinal implants. CNTF has been observed to slow photoreceptor loss for retinal degenerations using animal models with similar retinal phenotypes to MacTel (21). However, providing retinal treatment with CNTF is challenging as the blood-retina barrier might prevent the molecule to reach the photoreceptors. The retinal implants used by the Phase 1 and Phase 2 clinical trials contains genetically modified RPE cells that release CNTF directly in the retina (22). The implants were well tolerated and beneficial effect on the treatment was observed in both trials. The phase 2 clinical trial found that 24 months after the implant in the treated eyes there was a significant reduction in photoreceptor loss as well as a reduction in the expansion rate of the ellipsoid zone break, increased macular thickness, and increased reading speed ability (suggesting an improvement on the parafoveal scotoma). Given the success of these implants on the prevention of photoreceptors loss, a phase 3 clinical trial (23) is now underway.

Although this treatment is showing promise, it is important to recognise that the final end stage of the disease consists of neurosensory atrophy or fibrosis and translates into a definitive loss of vision with no possible treatment (2). For this reason, prevention, early diagnosis and early treatment are key for MacTel patients.

1.1.7 Inheritance and genetics

A positive family history has been frequently observed in MacTel cohorts (2). Many cases of siblings and twins concordant for the disease have been reported as well as several multi-generational families. This suggests that genetics play an important role in the occurrence of the disease. However, families with multiple MacTel patients also display incomplete penetrance, which is attributed to by the late age onset of the disease. In general, inheritance appears to fit a low penetrance dominant genetic model (24).

Given how important early treatment is for outcome, understanding the genetic basis to the disease has been considered crucial for future advancement in disease diagnosis. However, up until 2015 genetic studies, consisting of family-based analysis, had been unsuccessful, with no genes identified.

A publication in 2010 (25) performed candidate gene association studies using a list of candidate genes identified as related to retinal vasculature, involved in the transport of lutein in the retina, genes identified with microarray expression in mice with induced similarities of MacTel patients, genes identified through linkage studies and genes related to comorbidities of MacTel disease.

Importantly, no specific biological pathway was identified for this list of candidate genes. A study performed on 17 families (24) also identified a promising linkage peak on chromosome 1 at the locus 1q41-42 with a LOD score of 3.54.

Unfortunately, subsequent sequencing of candidate genes for causal mutations in this genomic region failed to identify any plausible pathogenic variants suggesting that the inheritance model may not be monogenic, but rather polygenic.

In this section, we have explored the clinical and biological, including genetic, understanding of MacTel disease available prior to the work developed f in this thesis and work by other researchers that have also advanced MacTel research. The next section will focus on different types of data and data analysis that formed the basis of the work conducted as part of this thesis to advance understanding of MacTel, in particular its genetic basis.

1.2 Dissecting the causes of MacTel disease in the age of 'omics data

In the first part of this chapter, we have described how MacTel is characterized by a multitude of phenotypes affecting different layers of the retina. Additionally, we have demonstrated that MacTel has a late onset of symptoms, comorbidity with heart disease, hypertension, elevated BMI, obesity and T2D. Lastly, a number of families have been observed where MacTel was diagnosed in multiple individuals, giving the first hint of a genetic contribution to the disease development. However, genetic linkage studies that aimed to detect causative variants did not identify any specific genetic alterations that commonly caused the disease.

Since the disease runs in families but no causal variant could be found in family studies, MacTel is likely to be a polygenic disease with multiple genetic factors affecting its aetiology. High prevalence of metabolic traits such as high BMI, T2D and chronic cardiovascular diseases is indicative of environmental or other non-genetic influences on disease aetiology, with the comorbid traits all known polygenic traits themselves. Lastly, the late onset might indicate that patients with this disease experience a complex continuous exposure to genetic and environmental factors that may result in disease manifestation.

All these considerations place MacTel in the realm of complex diseases whose aetiology often arises from intricate interactions of multiple factors. In the past 20 years, the realisation that many common diseases lie in this realm has highlighted the need to interrogate biological systems with multiple approaches, in a comprehensive manner (also referred to here as "systems biology"), in order to identify contributors to the underlying mechanism of these diseases. In the last decade substantial innovations and improvements in methods now permit system biology investigations that might help identify contributing factors in these complex diseases (26). The data that is derived from these unbiased approaches are referred to as 'omics data.

Each type of 'omics data is intended to measure a different "layer" from the basic genetic profile of each individual, which is not affected by any environmental factors, to the final disease phenotype which is a complex mixture of different genetic, phenotypic, and environmental influences. An adaptation from Hasin Y

et al., Genome Biol., 2017 (26) illustrates this continuum of the different 'omics data types represented as layers between the genetic background and the final disease outcome (**Figure 12**).

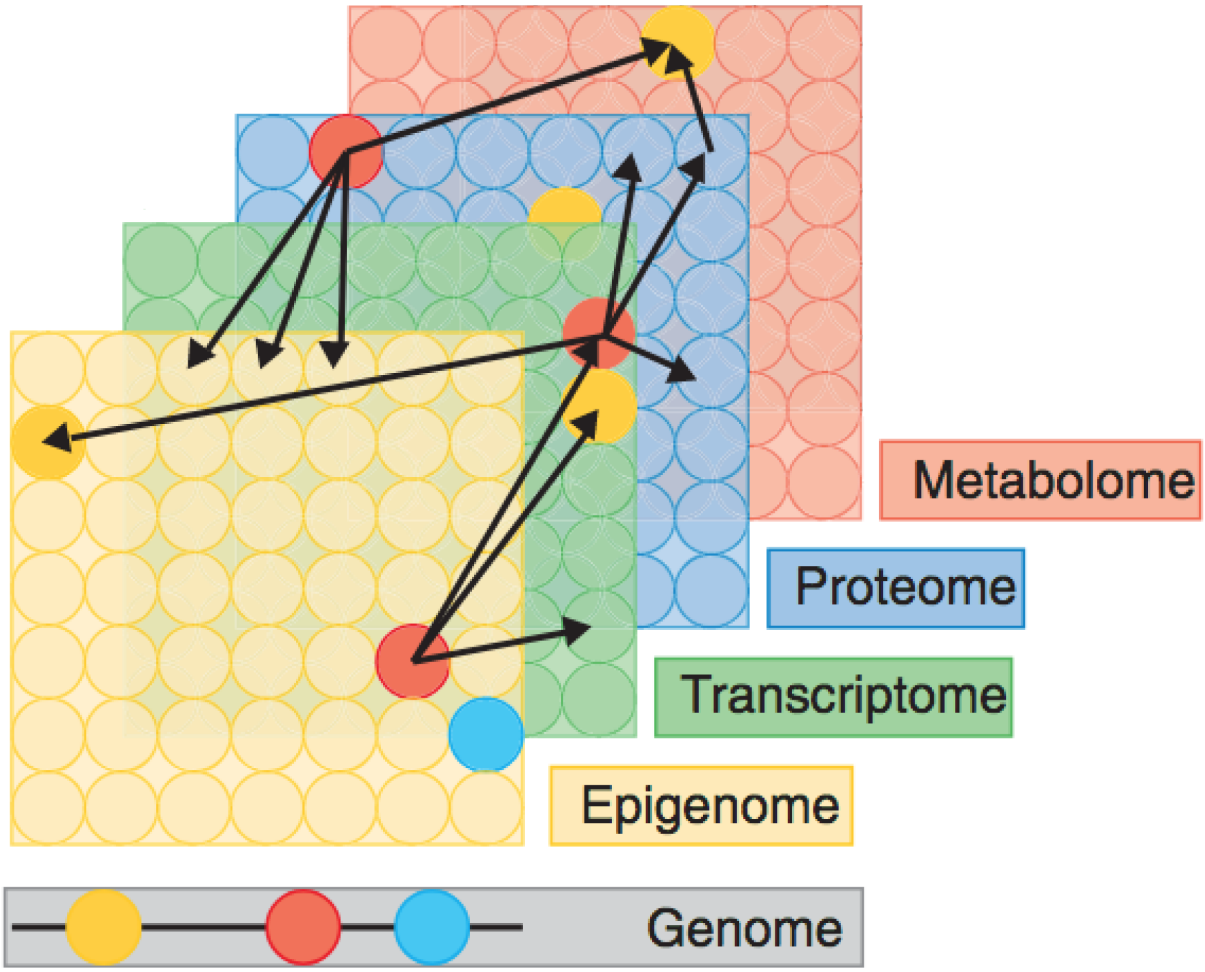


Figure 12: Visual representation of 'omics data on the space between the genetic background and final trait adapted from Hasin Y et al., Genome Biol., 2017 (26).

Each type of 'omics data measures different biological information, described below, which can be used to search for association with disease:

- *Genomics*: identify genetic variants in DNA

- *Epigenomics*: measure reversible modifications of DNA or DNA-associated proteins
- *Transcriptomics*: measure the level of RNA transcripts, which can be used to infer gene expression
- *Proteomics*: measure peptide abundance, modifications and interactions
- *Metabolomics*: measure the abundance of molecular compounds present in the body's metabolism
- *Phenomics*: measure the full set of phenotypes present in each individual

Piecing together multiple kinds of 'omics data makes it possible to unveil the complex web of genetic and environmental interactions that lie at the heart of the causal disease mechanism (26–28).

In the next section, we will describe some of the most commonly used types of 'omics data that were analysed in the research presented in this thesis.

1.2.1 Genomics data

The human genome consists of almost 6.5 billion bases. However, humans have more than 99.9% of their DNA in common (29). One type of common variation between individuals are called Single Nucleotide Polymorphisms or, more commonly called, SNPs. A visual representation of a SNP is provided in **Figure 1**3. SNPs are defined as substitutions at single positions in the genome that are shared by a considerable percentage of the population (eg. > 0.1%).

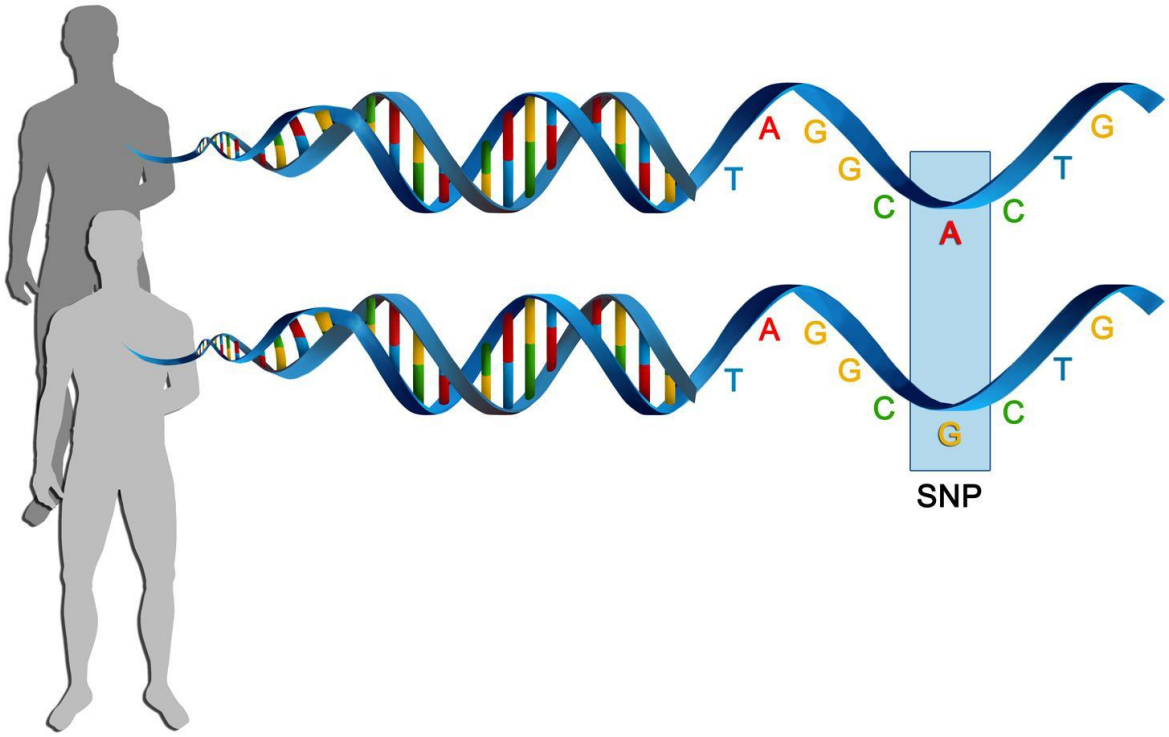


Figure 13: Visual representation of Single Nucleotide Polymorphisms (SNPs) adapted from (30).

SNPs are often classified by "rs identifiers" and each has a specific chromosome and base-pair position. For example, SNP rs715 is a common variant in chromosome 2 at the base-pair position 211,543,055 which represents a change from the common allele T to the less common allele C. As research and knowledge regarding the human genome progresses, different annotations for each SNP are released by governmental agencies and consortia such as the Genome Reference Consortium. Annotations define the SNP name, position, and allele change. Currently, the two commonly used annotations are the GRCh37 and the most recent GRCh38. **In this thesis, unless otherwise stated we will always refer to SNPs using their GRCh37 annotation.**

Depending on their position, SNPs can be defined as intragenic or intergenic, depending on whether the SNP is located within a gene or between genes. Intragenic SNPs can be further divided into exonic SNPs, located within the coding regions of a gene, and intronic SNPs, located between coding regions. Lastly, exonic SNPs can be divided into synonymous or non-synonymous SNPs. Synonymous SNPs do not change the amino acid during transcription whereas non-synonymous SNPs will, and as a consequence are usually considered more important since their presence is much more likely to lead to pathogenicity.

1.2.2 Metabolomics data

Metabolomics involves the measurement and profiling of chemical compounds in the human body called metabolites (31). As presented in **Figure 1**2 metabolites are much closer in "phenotype" to the final disease outcome than genetic data. In fact, metabolites are essential for the physiological functioning of the human metabolism. They can be measured with mass-spectrometry which detects the mass of different compounds in a sample and quantifies them.

There are two main approaches used to quantify metabolite abundances. The first is targeted metabolomics, which records and quantifies only specific metabolites, chosen by the user to investigate a specific hypothesis (32, 33). Targeted metabolomics can give very high sensitivity and selectivity and usually requires simpler and less intensive computational efforts (31). However, tagging only

specific metabolites requires the researchers to know *a priori* which metabolites might be associated with the trait of interest. In the case where not much is known about which metabolites might affect a trait, the second technique, referred to as untargeted metabolomics might be more appropriate. This technique relies on the measurements of hundreds of metabolites covering most of the metabolomics profile for each sample. Although less specific and more computationally intense, this approach guarantees the exploration of a wider spectrum of chemical components and their relative pathway, making untargeted metabolomics very attractive for the study of traits with limited *a priori* knowledge. For this reason, untargeted metabolomics studies have been used to "generate" hypotheses (33).

1.2.3 Phenomics data

Phenomics is an extremely broad term that has been previously used to describe any kind of 'omics data not strictly definable as genomics (27). Genomics data is intrinsically different from any other kind of 'omics data as the former can be generally considered as "fixed" or "stable" between different conditions. In comparison, phenomics data is "flexible" and "unstable" as most biological phenotypes can change depending on factors at the time of collection and generation of the data, such as: cell type, tissue origin, time, age and other external factors affecting the sample. This means that a complete characterisation of the phenome, which is always bound to a specific place and time (27), is impossible to achieve.

Given recent advancement and establishment of specific 'omics measurements such as metabolomics, transcriptomics, epigenomics etc the definition of the term phenomics has slightly sharpened, now often referring to phenotypic measurements closer to the disease.

In recent years, there has been a focus on the discovery of genetic variants associated with different traits and disease (34, 35). Phenotypic variability attributable to genetic contribution has been defined as *heritability* (defined and discussed in chapter 2). However, it has been observed that only a very small proportion of disease heritability can be explained by genetic variants currently detectable by genomic studies (36). Highly complex diseases often present with an extreme phenotypic heterogeneity leading to attempts to divide these into sub-disease groups, often using ad-hoc approaches, based on clinical observations which may not reflect true underlying biological heterogeneity, such as that due to different biological pathways being affected in different patients. Measuring phenomics information provides the opportunity to better group subjects phenotypically similar to each other; the assumption being that they will also be more genetically similar and thus allow simplification of the disease to increase understanding and potentially even lead to targeted therapeutic development and treatments. **In this thesis we refer to phenomics data as phenotypic data gathered through retinal imaging technologies. We provide extensive description of these technologies in Section 1.1.3.**

In this section, we have explored how different types of 'omics data might be collected to better explore the genetic and biological mechanism underlying MacTel. The next section will explore how 'omics data can be analysed and how different kinds of 'omics data can be integrated to construct a more comprehensive and hence better understanding of a trait.

1.3 Analysis and Integration of 'omics data

Analysis and integration of different 'omics data poses specific data analysis challenges and consequently substantial effort has been applied to develop analysis methods to make the most of these data. To put this in some perspective, in 2006 there were about 1000 publications that mentioned "data integration" in their abstract. Six years later, that number had doubled with around 2,300 publications addressing the same issue (28). In this chapter, we will provide an overview of the most common methodologies used to analyse 'omics data in this thesis. To clearly divide the methods used to analyse single omics data from methods integrating two different 'omics data from those integrating three, four etc, we will divide these methods into different sections (1.3.1 - 1.3.4) subclasses. We only introduce the methods that were used in this thesis, noting that a comprehensive review of this extensive field is not possible in this introduction.

1.3.1 Analysis of single 'omics data

1.3.1.1 Genomics + Disease = GWA studies

Genome-wide Association Studies, commonly referred to as GWAS, are experimental studies which aim to discover the association between genetic variants (usually SNPs) and a trait of interest within samples from specific genetic populations (34). The first wave of GWAS happened around the early 2000s with enormous expectations on the biological and translational results they would bring. However, by the end of the decade careful evaluation of GWAS results so far highlighted both achievements and pitfalls (37). Half a decade later, the doubts of GWAS utility peaked and addressing the misunderstanding about their efficacy became crucial (35). By the end of the 2010s, at the time of this thesis, the utility and applicability of GWAS were addressed again, and novel applications and new methods were discovered regularly (34).

When talking about GWAS it is important to address the fact that such studies commonly focus on identifying *common* genetic variants associated with a disease rather than *rare* variants. Common variants are usually defined as those variants that are "commonly" found in a specific population. Generally, SNPs have Minor Allele Frequency (MAF) greater than 0.01. GWAS focus on these variants because of the need for very large samples sizes required to identify enough affected subjects with the same disease to satisfy power requirements. In contrast, extremely rare variants are more likely to be unique to specific individuals and

can thus not be analysed with standard GWAS approaches. Identifying common variants in contrast to rare variants has the potential to elucidate a shared biological mechanism underpinning a specific disease. Interventions on such shared mechanisms should be beneficial for most of the affected individuals, in contrast to specific and personalised interventions.

It is now well understood that common variants surviving evolutionary selection, tend to have a reduced impact on disease risk compared to rare variants which are often observed to be extremely pathogenic. A visual representation from Manolio et al 2009 (36) of the effect sizes of genetic variants on general traits in relation to their allele frequency in the population is presented in **Figure 14**

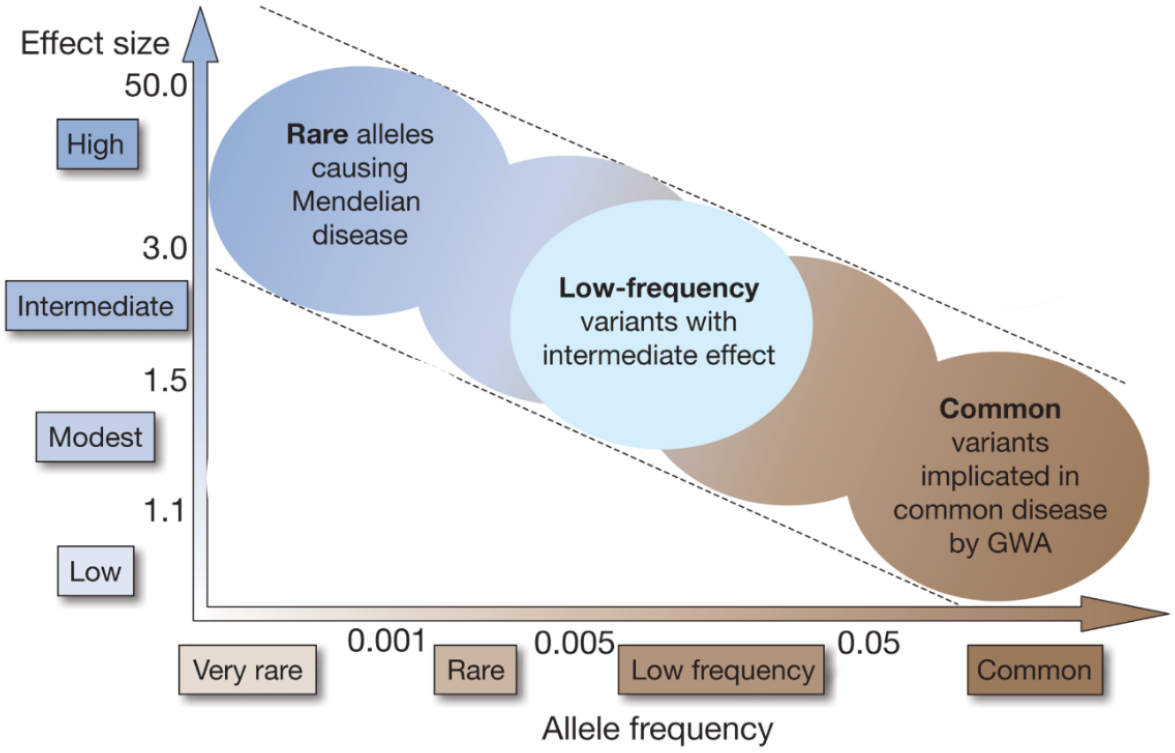


Figure 14: Genetic variants effect size in relation to minor allele frequency adapted from Manolio TA et al. Nature, 2009 (36)

Extensive review about GWAS methodologies are beyond the scope of this thesis and can be found elsewhere (38–41). However, a typical GWA study might be summarized as following (37):

1. Collect DNA samples from patients and healthy controls
2. Genotype SNPs using sequencing technologies
3. Clean genotype data by running quality control steps
4. Impute genotype data to infer SNPs not directly genotypes
5. Clean imputed data by running quality control steps
6. Perform GWAS analysis by testing for association of each SNP with the trait of interest by correcting for potential confounders
7. Identify genuine SNPs and genomic loci of interest
8. Interpret how such genomic loci might be involved in the disease

Since the goal of GWAS is to find common variants associated with a specific disease, such studies usually scan across the entire genome and independently test all SNPs for significant associations. However, typical GWAS test for millions of genotyped and imputed variants, resulting in an enormous multiple testing burden and an increase in the false discovery rate, where variants that might not be truly associated with a specific disease are significant at a nominal level. For this reason, is common practice to correct for false discovery rate using a strict p-value correction called Bonferroni correction. The standard correction internationally adhered to is $p<5e-08$. This threshold was determined based on

empirical linkage disequilibrium patterns, determined by projects such as the HAPMAP (42) and 1000 Genomes project (29).

As already mentioned, SNPs identified by GWAS studies often have small effect sizes on disease risk. Small effect sizes usually translate into p-values which would not survive correction for multiple testing necessary to robustly identify SNPs with a causative role in a disease. For this reason, very large sample sizes are usually required to perform GWA studies. Such sample sizes also usually reflect the rarity of a disease. Rare diseases may require smaller sample sizes as the number of genetic variants affecting such traits might be small, with larger effect sizes.Other and more common traits might be contributed to by hundreds of variants and can be considered as extreme polygenic traits. GWA studies on such traits often require samples sizes in the order of hundreds of thousands of individuals. Recent examples of such mega GWAS include studies into height (43) and education attainment (44).

In the specific setting of this thesis, MacTel is a relatively rare disease for which no single rare genetic variant could be robustly identified using family-based analysis approaches. Because of this, we might expect that common genetic variants might play an important role in the disease and that moderate sample size would be sufficient to reliably identify a subset of important SNPs which might help elucidate the disease. Results of the first GWAS ever performed on MacTel will be presented in Chapter 2.

1.3.1.2 Metabolomics + Disease = Metabolomics Study

Untargeted metabolomics studies focus on measuring metabolic abundances across the entire metabolome spectrum in both healthy subjects and disease patients to subsequently identify metabolites which might be involved in the disease. Metabolites found to be associated with a trait by untargeted metabolomics studies are often called metabolic "biomarkers". Unlike genetic variants, biomarkers are closer to the observed disease outcome and are expected to have bigger effect sizes, requiring a smaller sample size compared to typical GWA studies.

However, metabolomics studies do not come without challenges. Firstly, as mentioned before, metabolomics as well as all other phenomics profile cannot be uniquely characterised and are reflective of a specific time and place. For example, the metabolic profile of the same individual would be different if extracted from different biological samples such as blood or urine. The same profile would be additionally different if the same sample was to be extracted before or after fasting. This profile will also differ if the sample was taken during childhood or during adulthood. Moreover, sample storage conditions, machines and techniques used to measure metabolic abundances and even human interference in sample processing can alter specific metabolic profiles inducing unwanted variation (45, 46) that might hide genuine biological signal. As for GWAS, intensive research effort has been invested in developing methods aimed at overcoming such

complications. (45–48). A general untargeted metabolomics study might be summarised as follows:

1. Collect biological samples and measure metabolomics abundances across the entire metabolome spectrum.
2. Perform QC steps, identify batch effect and confounders, and normalise the data to account for technical variation
3. Identify metabolic biomarkers by performing differential abundance between cases and controls for each metabolite
4. Explore whether the metabolic biomarkers are indicative of a specific metabolic pathway or group

The identification of metabolic biomarkers has been historically the main goal of metabolomics studies (31). However, metabolomics has seen a rapid expansion of research goals in recent years and advances in both technical and instrumental performance, as well as the development of bioinformatics tools and publicly available databases. This has helped to move from the simple biomarker discovery to the more complex analysis of interconnectivity in the metabolome and in metabolomic pathways.

As previously mentioned, we can expect MacTel disease to be affected by complex genetic factors as well as and non-genetic factors which will affect the 'omics between DNA and eventual disease. Metabolomics offers an opportunity to explore this intermediate 'omics space and indeed may better capture the mechanism

behind the disease aetiology. Results of a standard semi-targeted metabolomics analysis are presented in Chapter 2 while the results of a comprehensive untargeted metabolomics study are provided in Chapter 4.

1.3.2 Joint 'omics data integration

1.3.2.1 Genomics + Transcriptomics = eQTLs Study

One of the most important questions usually asked in GWA studies is the functional mechanism that might be behind the association between genetic variation and a trait. One of the first steps towards answering this question is to investigate which genes' expression might be affected by the SNPs of interest in the association peak. This question might be trivial for intragenic SNPs which might even be located within the coding region of a gene. However, for intergenic SNPs, which account for the majority of GWAS signals (37) the answer to these questions is often far from straightforward. In fact, it has been reported that only ~1/3 of eQTL signals it is the nearest gene to the SNP of interest (34). A frequently used approach, is expression Quantitative Trait Loci (eQTL) analysis (49), with different methods having been proposed to perform such analysis (50). Typically, eQTL studies use linear regression to test whether changes in SNPs can predict the level of gene expression of candidate genes whose expression can be summarised with gene expression levels determined from microarray, or more

frequently now, RNA-seq data. However, it is important to remember that gene expression is highly variable, tissue-specific and noisy. For this reason, in 2008 the Genotype-Tissue Expression (GTEx) project was proposed, with the intention to characterise gene expression across different tissues and cell lines of the human body (51). A visual representation of the tissues and cell-lines explored in GTEx is presented in **Figure 15**.

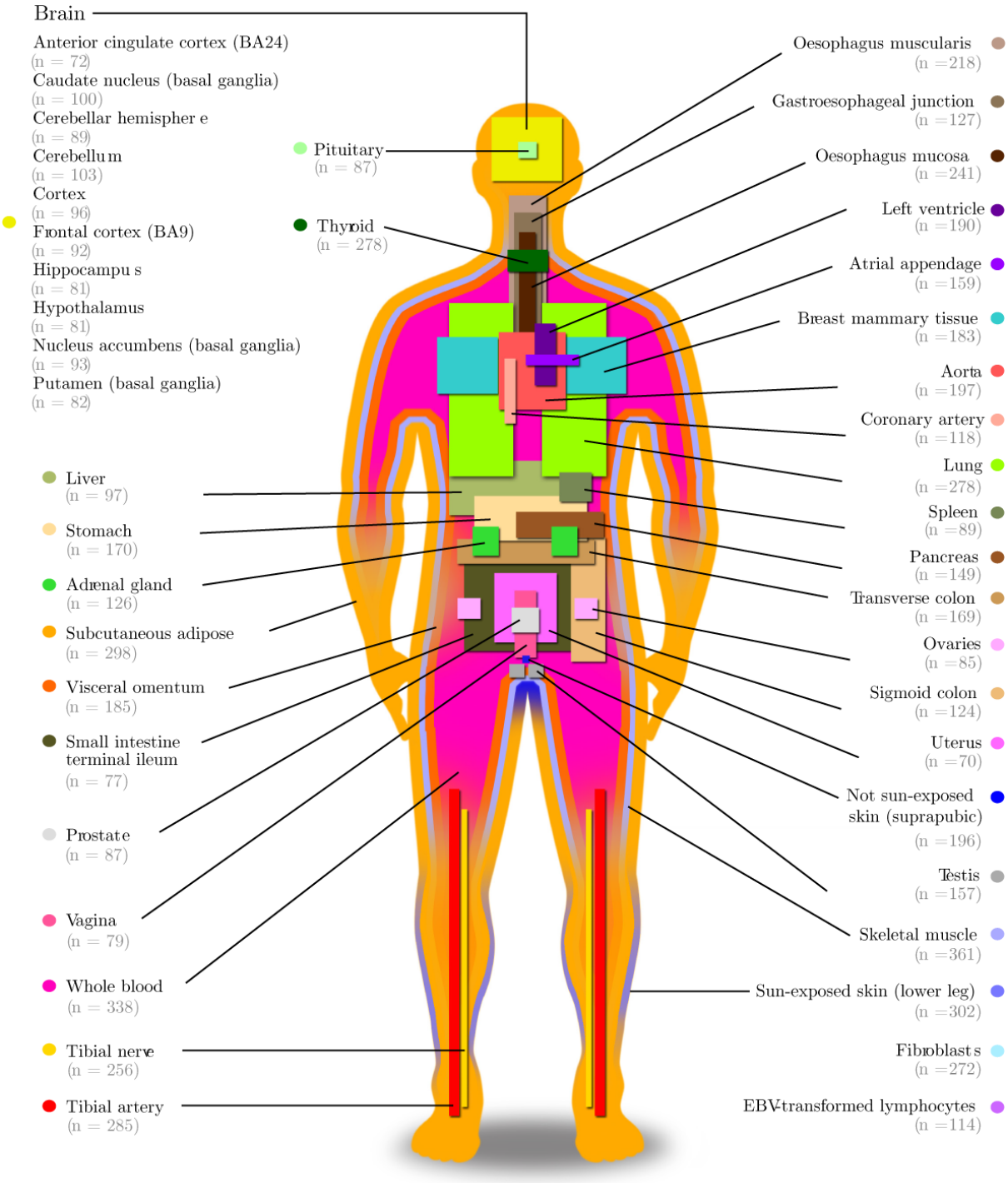


Figure 15: Visual representation of all tissues gene expression and relative sample size explore by the GTEx consortium. Note the absence of any eye tissue, including retina. Image adapted from Battle A et al., Nature Publishing Group, 2017 (51).

Apart from collecting gene expression from different tissues, the GTEx project also explored the effect that common genetic variants like SNPs might have on gene expression, stratifying for tissues and cell lines. This project has greatly contributed to our understanding of both the transcriptome and eQTL effects across the human body and provides an invaluable tool for exploring the functional effect of identified SNPs such as those found to be associated with traits in GWAS studies.

An important distinction needs to be made when describing eQTLs. SNPs affecting gene expression through a direct pathway are called cis-eQTLs while SNPs which affect gene expression through a distant pathway are called trans-eQTL (49). However, the specific biological pathway through which SNPs affect genes is mostly unknown. Usually, SNPs located within the transcription boundaries of a gene or located near in a region near a gene, which might contain a promoter or an enhancer for that particular gene, are assumed to affect the expression of such genes directly, while distant SNPs are assumed to have an indirect effect. For this reason, SNPs affecting the expression of neighbouring genes (usually less than 500Kbp or 1Mbp apart) are called cis-eQTL, while those affecting more distant genes are called trans eQTL (49).

Testing for eQTLs usually involves testing associations between millions of SNPs with thousands of genes. This, in turn, results in an enormous multiple testing burden which is known to increase false positive discovery. For this reason, testing is usually performed separately between cis-eQTLs and trans-eQTL. Because of

this approach, cis-eQTL are in general easier to find than trans-eQTL. Firstly, the amount of testing required for cis-eQTL is usually much smaller than for trans-eQTL. Secondly, the effect size of cis-eQTL tends to be in general larger than for trans-eQTL, with evidence of intergenic cis-regulation of SNPs on genes to lie between 800Kbp and 1.3Mbp away from the genes' transcription start site (51).

1.3.2.2 Genomics + Metabolomics = mQTL Study

Similarly to the concept of eQTL studies, researchers have been interested in common genetic variants associated with any measurable quantitative trait. Such studies have been defined generally as Quantitative Trait Loci studies (QTLs). Understanding the genetic background of metabolic regulations has the potential to inform on the mechanisms and pathways involved in the biosynthesis of different metabolites as well as genomic factors that might be involved in the modulation of metabolic profiles. However, unlike in the eQTLs setting, there is no prior information that can be assumed to divide trans from cis effects of specific SNPs on metabolomics abundance. For this reason, metabolomics Quantitative Trait Loci (mQTL) studies are often performed as several GWAS, where metabolomics datasets are combined with genomics information and a GWA study is performed on the abundance of each metabolite. Only recently, datasets of a sufficiently large size, which can overcome the tremendous amount of multiple testing and the problem of relatively small effect sizes of SNPs on metabolomics abundances have become available. However, the few studies performed so far (52) have already created exciting resources, such as the metabolomics GWAS server

(53), that can be interrogated to discover whether specific SNPs of interest are related to the abundance of specific metabolites.

1.3.3 Three-way ‘omics integration

1.3.3.1 QTLs + Genomics + Disease = Mendelian Randomization

By combining information from QTL studies for specific traits with genomic information from other traits it is possible to perform Mendelian Randomization (MR). MR is a relatively new and growing field of research in genetics gaining prominence in the last decade. It aims to detect unbiased **causal** effects and possibly estimate their magnitude (54).

Before explaining how to perform MR through data integration, it is important to first address causality. To do this we will make use of the work of (54–56). Imagine a child that wants to play in a park, but the child is sick and needs to stay home and take antibiotics. In this example the child’s health condition, taking antibiotics, and playing in the park are all associated with each other but the health condition of the child that is causing both circumstances: taking antibiotics, and not playing in the park. Hence, in general, a child taking antibiotics and not playing in the park are associated to each other (co-occur), but there is no causation between the two. Causality has been historically represented by using

Directional Acyclic Graphs (DAGs) and theoretical considerations on DAGs has been deeply explored (57). The above example can be represented as in **Figure 16**.

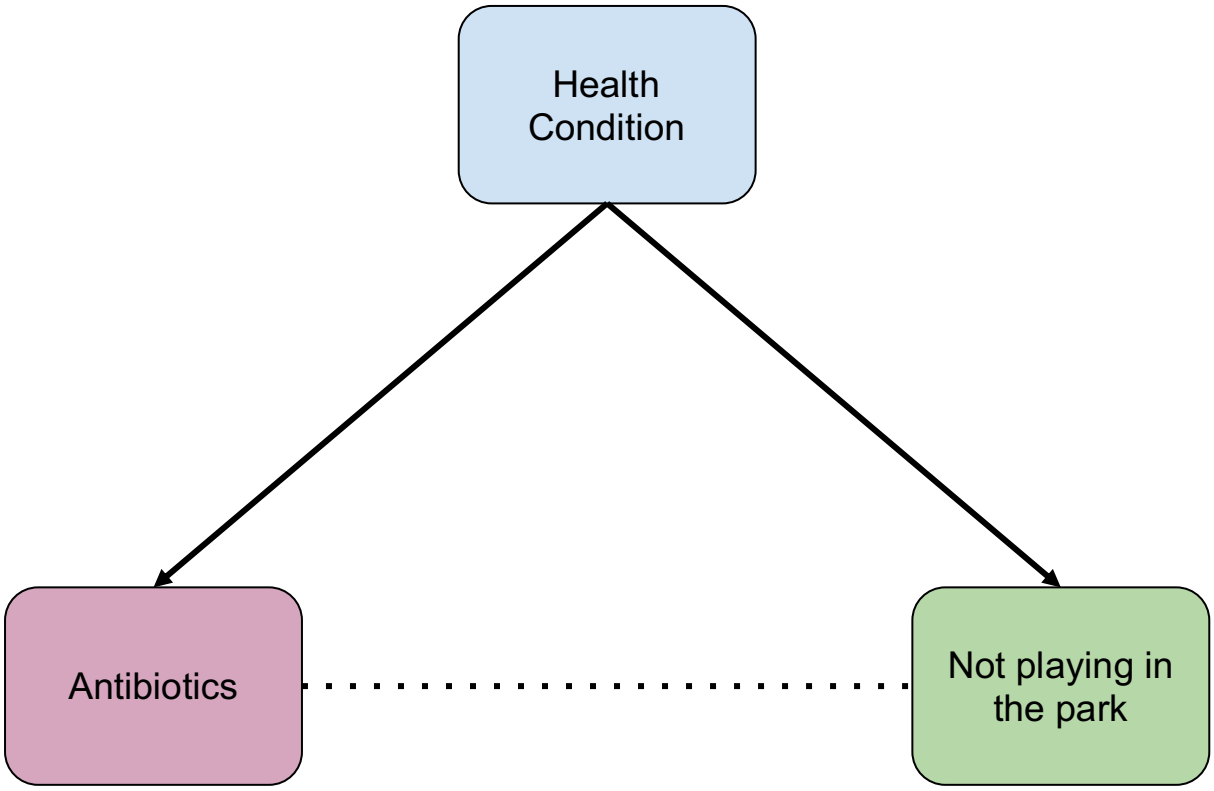


Figure 16: Directional Acyclic Graphs demonstrating association (dotted line) and causation (solid arrows) between three different phenomenons. Dotted lines represent associations while solid arrows represent causation.

In **Figure 16** we see that causal associations are presented as solid arrows while associations are presented as dotted lines. As we see in this example the health condition is connected with both antibiotics intake and not playing in the park, and this connection arises from a *causal effect* of the health condition on the two. Antibiotics intake and not playing in the park are also associated but this association is not due to any causative effect.

Association not due to causation may arise from several different scenarios and in this case, arises from a phenomenon known as *confounding*. In fact, in this case, the confounding effect is the health condition which, if not taken into account, will make antibiotics and not playing in the park associated with each other.

Another example of association not due to causation is known as reverse causation. In our example, antibiotics intake and the health condition are clearly associated. If a naive researcher observed that everyone who takes antibiotics also has a health condition, this researcher concludes that antibiotics are associated with a health condition. By not knowing why this is the case, the naive researcher may think that antibiotics are causing the health condition where of course, the opposite is true. So, in this case, the association between antibiotics and the health condition arises from a *reverse-causation* of the health condition on antibiotics as shown in **Figure 17**.

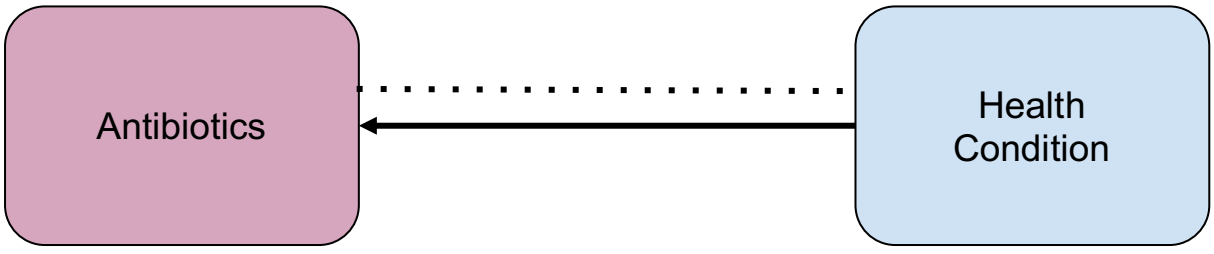


Figure 17: Directional Acyclic Graphs demonstrating reverse causation. Dotted lines represent associations while solid arrows represent causation.

To illustrate the differences between association and causation, we have used so far a very naive example for which anyone would be able to distinguish between

the two. However, most research studies focus on associations for which the directionality of the causative effect is not known (54–56). Imagine we studied the effect of blood levels of vitamin C on cancer risk. It might be possible to observe that people with cancer have low vitamin C. This might suggest that having low vitamin C levels may predispose to increase cancer risk (direct causation). However, can we really claim this? Maybe cancer cells use vitamin C in order to duplicate faster (inverse causation). Also, we know that people with high levels of vitamin C are likely to have a healthy diet, which itself is common among people living a healthy lifestyle, who again are likely to avoid many of the common risk factors for cancer like smoking (confounding). Also, people suffering from cancer may modify their diet due to illness resulting in a depletion of vitamin C (indirect inverse causation). Even in this relatively simple example is it easy to recognise that the association between vitamin C and cancer risk is anything but obviously causal. Classic epidemiological studies have generated thousands of replicated associations between traits. However, these studies are often unable to claim a clear causative effect.

Causation has been historically tested by performing Randomized Clinical Trials (RCTs) (54–56). These studies involve *randomly* exposing individuals lacking any trait of interest (eg. totally healthy individuals) to the exposure of interest (eg. Vitamin C), then waiting for a specific amount of time and assessing how many of these individuals develop the final trait of interest (eg. cancer). It is easy to recognise why RCTs are protected from both confounding and reverse causation

as people are assigned to receive the exposure and the outcome completely at random, preventing confounding.

There have been many examples of associations repeatedly reported by epidemiological studies, which have not been confirmed as causal associations by RCTs. Some of these include the association between the antioxidant vitamin ß-carotene and smoking-related cancers, or the association between Vitamin E and coronary heart disease (55).

Clinical trials are not possible for many traits for ethical and feasibility reasons (eg. cigarette smoke and cancer causation).

Fortunately, genetics provides an indirect RCT in that every single human being is randomly assigned to a combination of genetic factors. Genetic factors contribute to thousands of different exposures and are largely unaffected by any external influence. Because of this, we can think of such genetic factors as "life-long exposure treatments". In our example, we imagine that people with specific genetic factors, predisposing them to smoke, are the equivalent of a randomized group of people in a trial exposed to cigarette smoke. If we observe that people carrying these genetic factors also develop cancer later in life, we can claim that smoking might actually have a causal effect on cancer risk. This analogy between RCTs and MR has been beautifully explained in a YouTube video by Prof. George David Smith https://www.youtube.com/watch?v=LoTgfGotaQ4.

A more technical explanation of MR comes from *Instrumental Variables* (IVs) which have been historically used to test for causality (54). This method states that IVs capturing a specific exposure are associated with a trait only in the presence of direct causation. Theoretically, IVs satisfy the following properties:

1) The IV needs to be reliably associated with the exposure of interest
2) The IV needs to be independent of unobserved confounders that influence exposure or outcome.
3) The IV needs to be associated with the outcome only through its effect on the exposure

Genetic variants such as SNPs satisfy almost all these properties. In fact, reliable associations between SNPs and traits are obtainable through QTL studies and a SNP is independent of most confounding factors that may influence either outcome or exposure. However, there is no unique way of testing whether a specific SNP affecting an exposure of interest also affects an outcome only through the initial effect on the exposure. In fact, genetic variants such as SNPs are often observed to have pleiotropic effects which are considered to be the main limitation of Mendelian randomization. Consider the possible scenarios presented in **Figure 18**.

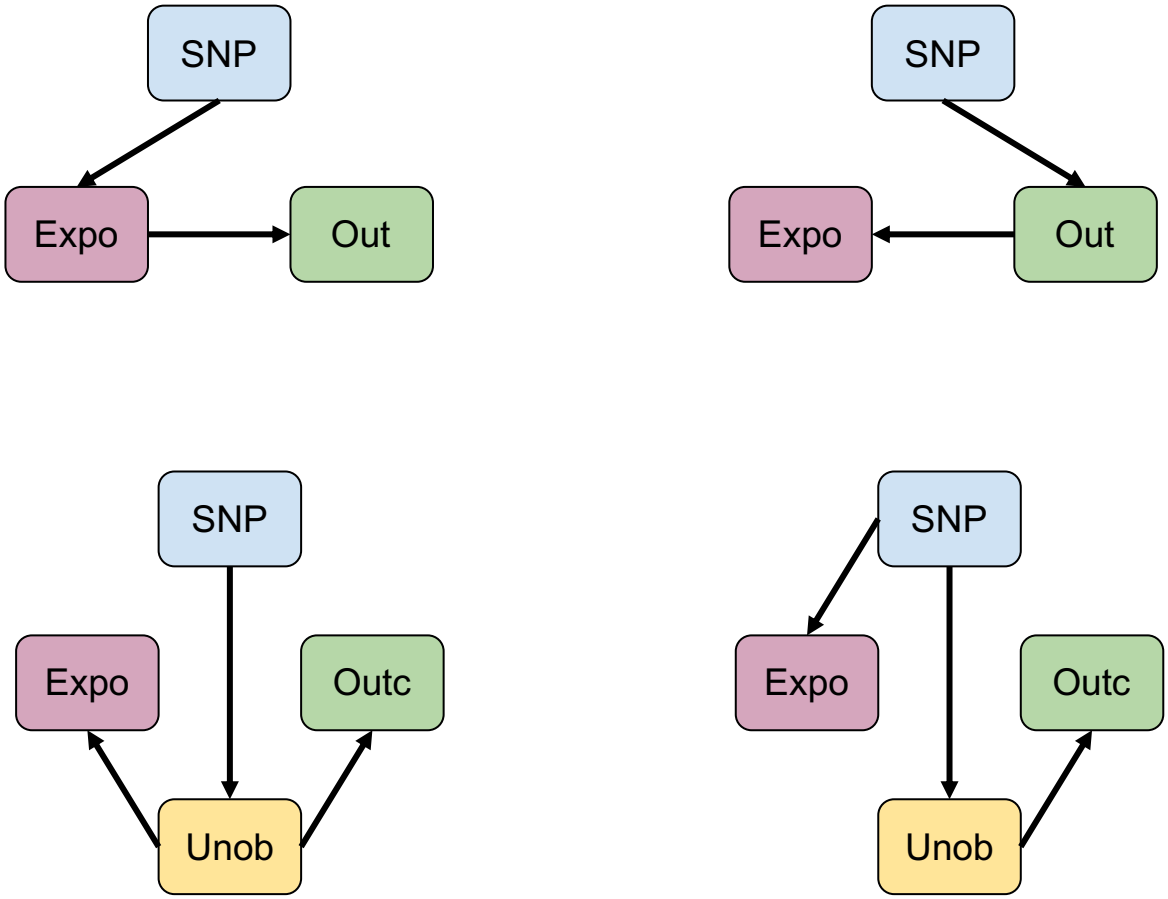


Figure 18: DAGs demonstrating of different scenarios. A Direct causation scenario. B Inverse causation scenario. C Pleiotropy type 1 scenario. D Pleiotropy type 2 scenario.

In **Figure 18**, the SNP will be associated with the exposure as well as with the outcome in all four scenarios, and because of this, the exposure would be considered as causative for the outcome. However, only the scenario presented in **Figure 18** A is truly indicative of a causative effect of the exposure on the outcome. A strategy has been implemented in MR techniques to overcome misdetection of causative effects in situations B, C and D. This strategy first extracts multiple independent SNPs associated with the exposure, then tests them all for

association with the outcome. If most of them are also associated with the outcome then the causative relationship is deemed to be genuine. We can, in fact, expect that not all SNPs associated with exposure present the same pleiotropic effects. Imagine if the exposure is not causative for the outcome. In this case, we expect that even if a single SNP is associated with both exposure and outcome because of a pleiotropic effect or inverse causality, all the other SNPs associated with the exposure not presenting the same pleiotropic effect should not be associated with the outcome. Hence the combination of all these SNP associations should suggest a non-causal relationship between exposure and outcome.

There are many ways of implementing MR between an exposure of interest and a specific outcome. Most methodologies can be grouped depending on the data available. There are methodologies developed to test MR on data where all three types of 'omics data are available for the same individuals i.e. genomics, phenomics (exposure), and outcome. Other MR method apply when the phenomics information on the exposure is missing and only the general QTL effects, capturing SNP effects on exposure, are available, as well as genomics data and outcome. A third MR method applies when the genomics data is missing as well and the available information only comes from QTL effects and GWAS results from SNP effects on the outcome. An extensive review of these implementations is beyond the scope of this thesis and can be found elsewhere (54–56).

MR allows investigation of different questions which might be more complicated than a simple direct causative effect. A methodology of particular interest to our

study is called hypothesis-free MR that can be used to test for causal association between exposures and outcomes which has never been explored before (eg. expression of specific genes or metabolites to disease risk). One can use hypothesis-free MR to test for causative associations between any trait that is measurable using genomics data and MacTel disease. This translates to the invaluable opportunity of testing for factors that were never directly measured in the MacTel cohorts.

1.3.4 Multi trait genomics integration and the concept of genetic correlation

By integrating summary statistics from GWAS performed on several traits, one can explore the concept of genetic correlation between traits. Genetic correlation has been defined as the proportion of variance that traits share due to genetic causes (58). Historically, genetic correlations have been tested by measuring different phenotypes across different families and estimate the frequency of such phenotypes between them (59). However, such studies are often costly and require researchers to investigate multiple phenotypes on the same individuals.

With the advent of GWAS data, methodologies have been developed to explore genetic correlation using SNP data. In fact, MR can be seen as a particular method of testing for genetic correlation. However, as mentioned earlier MR requires SNPs that act as IVs which are usually those presenting an association with the trait of interest that reaches genome-wide significance. Because of this, MR can

only be performed for traits where such SNPs are available. However, this is not possible for a large number of traits where there are currently no or not enough genome-wide significant SNPs from GWAS results, perhaps due to as yet to low sample size and thus insufficiently powered studies,

For this reason, tools have been developed to estimate genetic correlations using all SNPs instead of only the significant ones. Recently, a tool that is able to estimate genetic correlations between traits using genome-wide summary statistics data called LD-score regression, has been published (59). This method not only reliably estimates genetic correlation between traits from publicly available datasets, but is also able to take into account that sample overlap might be present between two different GWAS summary statistics.

This study highlights the discovery potential that multi-omics data integration is able to achieve. By using such methods one can not only explore connections between traits but also prioritise biological pathways that, if shared between two genetically correlated traits, might be causal.

Lastly, a final example of multi-omics data integrations initiated by the GWAS community, also based on genetic correlations, is a tool called MTAG (60). This method aims to discover new GWAS hits by comparing GWAS results of genetically correlated traits. In short, this tool firstly checks for genetic correlation between the traits using the aforementioned LD score regression, then it prioritises SNPs showing shared effects on both traits. For example, if a SNP has

a positive genome-wide significant effect on trait 1 and has a suggestive positive effect on trait 2 and these two traits are positively genetically correlated, then this method will highly prioritise such a SNP also for trait 2, even though it did not originally reach genome-wide significance. This method becomes extremely powerful when we need to detected SNPs reliably associated with traits for which GWAS currently holds small discovery power (perhaps due to the difficulty of collecting a large enough study cohort). MTAG is able to, in part, address such problems and potentially identify SNPs that might have been missed by a specific GWAS because of this replication problem.

Using MTAG the authors were able to show that combining three genetically correlated neurological disorders together resulted in an increase in discovery power from the initial 32, 9, and 13 genome-wide significant loci to 64, 37, and 49 genome-wide significant loci respectively on the three traits.

To conclude, we can expect that combining the genomic information for several traits and MacTel using MTAG, might lead to the discovery of new potential genomic loci that might have been missed due to the small sample sizes that such a rare disease is limited to.

1.4 Conclusions

In conclusion, in this chapter, we have explored MacTel disease, the different types of 'omics data that might be collected to study such a disease as well as the different methodologies that might be used to analyse and integrate such multi-

omics data. The next four chapters will present studies performed on MacTel using the aforementioned data types and methodologies.

2 Contributions to Scerri et al. Genome-wide analyses identify common variants associated with macular telangiectasia type 2

2.1 Introduction

This chapter will describe the contributions that RB brought to the study "Genome-wide analyses identify common variants associated with macular telangiectasia type 2" by Scerri et al published in Nature Genetics in 2017 (61). The chapter will firstly describe the study and the results found, without the contribution of RB. Then, an introduction to the main framework of the analysis performed by RB will be given. Methods used to perform such analyses, as well as results and discussion arising from this study will then be presented later in the same chapter. The Scerri et al publication can be found in the supplementary appendix.

2.1.1 The MacTel GWA Study

In 2017 we published the first GWA study on MacTel using SNP chip genotyping performed using the SNP chip Illumina Infinium Omni5Exome-4 on a cohort of 476 MacTel patients and 1,733 controls to discover potential genetic loci involved in MacTel (61). This chip was able to directly genotype around 2.5 million SNPs.

Further imputation using the 1000 genome project reference panel returned almost 85 million SNPs in total. SNPs were then discarded based on low imputation quality, extremely low Minor Allele Frequency (MAF) and Hardy-Weinberg equilibrium, leaving around 6 million SNPs for disease association testing. After performing the GWAS analysis, which tested with logistic regression models the association between each SNP and the disease while correcting for sex and the first two principal components, we identified 28 suggestive ($P < 5e^{-05}$) loci involved in the disease, three of which reached genome-wide significance ($P < 5e^{-08}$). A Manhattan plot of the MacTel GWAS results is presented in **Figure 19**.

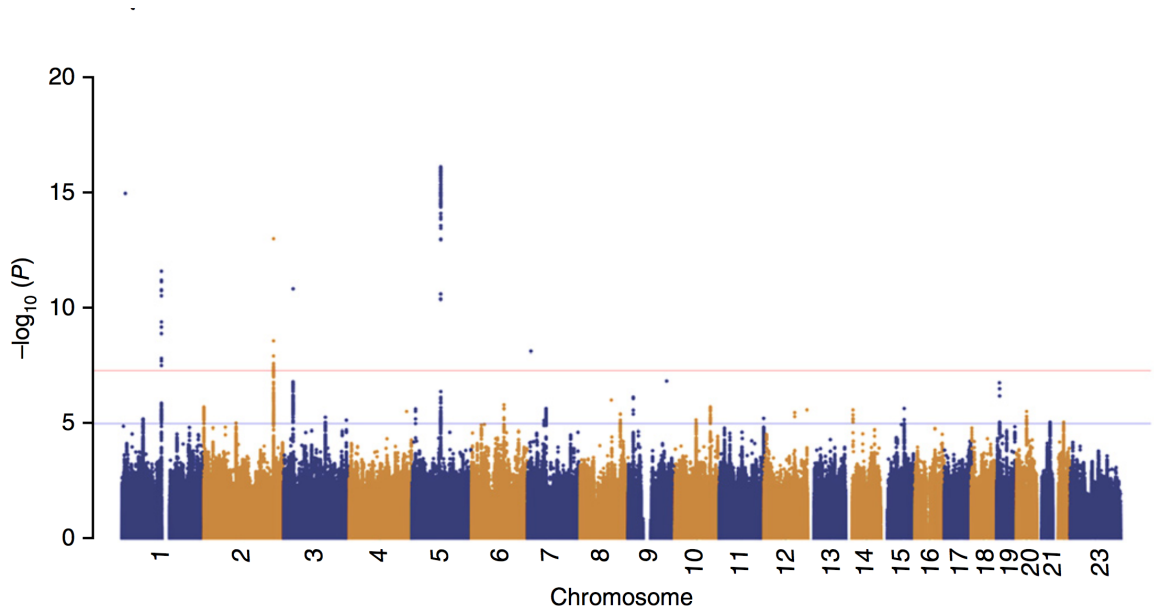


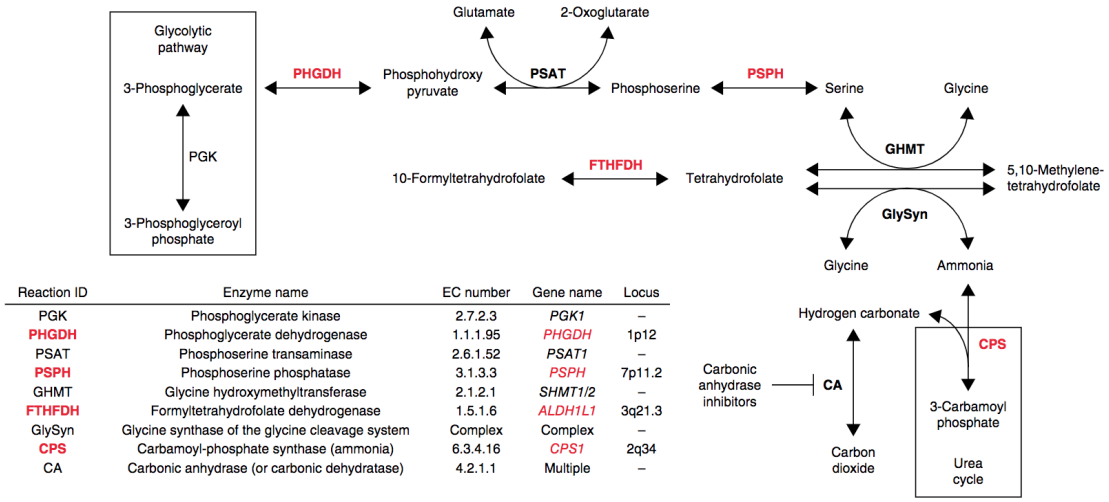
Figure 19: Manhattan plot displaying significant hit in MacTel GWAS as taken from Scerri et al, Nat Genet 2017 (61).

Given the moderate sample size of this study, only three loci reached genome-wide significance. These were tagged by SNP rs73171800 in locus 5q14.3, ($p=7.74e^{-17}$), SNP rs715 in locus 2q34 ($p= 9.97e^{-14}$) and SNP rs477992 in locus 1p12, (p =

$2.60e^{-12}$). We subsequently replicated associations at these three loci in an independent cohort of 172 cases and 1,134 controls. In the same cohort, we also replicated a further two loci that had not reached genome-wide significance in the initial cohort, but which were considered mechanistically related to the SNPs located in locus 1p12 and 2q34. These were SNP rs9820286, located in locus 3q21.3 ($p=5.38e^{-06}$), and SNP rs4948102 in locus 7p11.2 ($p=2.25e^{-06}$).

The SNPs identified in locus 5q14.3 had been previously associated with retinal venular and arteriolar calibre (62, 63). Subsequent work has shown that ablation of the homologous chromosomal region in zebrafish severely disrupts retinal development (64). The other four loci 1p12, 2q34, 3q21.3, and 7p11.2 had been previously associated with the abundance of glycine and serine in blood serum (52, 65, 66). Specifically, SNPs in locus 3q21.3 were in close proximity, but not in linkage disequilibrium with the SNPs found to affect glycine and serine ratio by Xie W, et al. Diabetes, 2013 (65). Additionally, loci 2q34, 1p12, and 7p11.2 all contained, or were in close proximity to, genes catalysing enzymes involved in glycine and serine biosynthesis. SNP rs715 in locus 2q34 was located in the 3' UTR region of gene *CPS1* which catalyses the Carbamoyl-phosphate synthase (CPS) enzyme; SNP rs477992 in locus 1p12 was an intronic SNP on the *PHGDH* gene catalysing Phosphoglycerate dehydrogenase (PHGDH) enzyme; SNP rs4948102 in locus 7p11.2 was an intronic SNP in the gene *PSPH,* catalysing the Phosphoserine phosphatase (PSPH) enzyme. Lastly, rs9820286 in locus 3q21.3 was located nearby to the gene *ALDH1L1,* which catalyses the Formyltetrahydrofolate dehydrogenase (FTHFDH) enzyme. All the aforementioned enzymes were found

to be closely connected to serine and glycine biosynthesis pathway as shown in **Figure 20**.



Reaction ID	Enzyme name	EC number	Gene name	Locus
PGK	Phosphoglycerate kinase	2.7.2.3	*PGK1*	–
PHGDH	Phosphoglycerate dehydrogenase	1.1.1.95	*PHGDH*	1p12
PSAT	Phosphoserine transaminase	2.6.1.52	*PSAT1*	–
PSPH	Phosphoserine phosphatase	3.1.3.3	*PSPH*	7p11.2
GHMT	Glycine hydroxymethyltransferase	2.1.2.1	*SHMT1/2*	–
FTHFDH	Formyltetrahydrofolate dehydrogenase	1.5.1.6	*ALDH1L1*	3q21.3
GlySyn	Glycine synthase of the glycine cleavage system	Complex	Complex	–
CPS	Carbamoyl-phosphate synthase (ammonia)	6.3.4.16	*CPS1*	2q34
CA	Carbonic anhydrase (or carbonic dehydratase)	4.2.1.1	Multiple	–

Figure 20: Proposed connections between candidate genes and the glycine-serine pathway as shown in Scerri et al, Nat Genet 2017 (61).

2.1.2 Considerations arising from the initial results and chapter aims

The results arising from the initial GWAS analysis and the genes implicated, clearly highlight a potential involvement of glycine and serine biosynthesis on MacTel risk. As part of the publication we also investigated several additional, important questions for MacTel genetic aetiology. Firstly, we investigated the genetic contribution to the disease or in other words, how much of MacTel risk could be explained by common genetic variations compared to other external factors. Secondly, we explored more deeply whether common genetic variants may be used to reliably predict disease risk or presence. Thirdly, by mining publicly available datasets like GTEx exploring relationships between SNPs and the

expression of genes we pinpointed multiple candidate genes for which expression might be affected by significant or suggestive significant SNPs. Lastly, we performed a metabolomics analysis with an independent set of MacTel patients and healthy controls to validate the involvement of glycine and serine on MacTel.

RB performed these four analyses which formed part of the MacTel GWAS publication (61). The following section will introduce and present the results of these four investigations.

2.1.3 "How much" MacTel can be explained by genetics: The heritability concept

A common question asked in GWA studies is how much of the phenotypic variability can be explained by genetics, or in simpler words, how important is the genetic component for disease risk?

This question is usually explored with the concept of heritability. Heritability has two main definitions:

- The **narrow-sense heritability** defines heritability as the proportion of total phenotypic variation explained by additive genetic factors effects.
- The **broad-sense heritability** defines heritability as the total contribution of genes on the phenotypic variation.

For our publication, we were mainly interested in the narrow-sense heritability, since multiplicative gene-gene and gene-environment interactions are usually

hard to identify and measure, especially with SNP genotyping data. The heritability of a trait is usually estimated from pedigrees and twin studies. However the rarity of MacTel, it's late-age onset and the difficulties in diagnosis provide too few families to investigate this using these approaches, and very few twins with MacTel have been reported (2). These heritability estimates, vary between different traits. For example, height is an intensely investigated, highly polygenic trait, for which heritability has been estimated to be around $h^2 = 0.8$ (67, 68). Recently, methods have been developed to estimate heritability using SNP data (68, 69). By 2010, GWA studies had discovered over 40 common variants associated with height. However, these variants together only explain $h^2_{known} = 0.05$ of the phenotypic variance i.e. they explain $h^2_{explained} = 6.26\%$ (=0.05/0.8) of the estimated total heritability. This proportion was called "Explained Heritability" by Zuk et al in 2012 (70). Manolio et al 2009 (36), coins this problem as one of "Missing Heritability". Missing heritability is defined as the proportion of estimated heritability not explained by common variants, $h^2_{miss} = 1 - h^2_{explained}$. Several hypotheses have been formulated for this phenomenon. The general consensus is that most of the missing heritability might be explained by SNPs that are below the GWAS significance threshold (71), or by low-frequency variants, too rare in any given population, whose effect will not be captured by standard GWA studies as previously shown in **Figure 12**. Other sources contributing to this missing heritability are also copy number variations (CNVs) and short-tandem repeats (STRs). As demonstrated(67), the heritability for height estimated by using all SNPs (after some correction regarding the MAF of the causal SNPs) reached 80%. An interesting argument raised is that genetic interactions create

"Phantom Heritability" (70). No heritability estimate had ever been calculated for MacTel hence we were interested in estimating not only the total MacTel heritability (h^2), but also the known (h^2_{known}) and explained ($h^2_{explained}$) heritability, estimated using the five replicated loci, as well as the proportion of missing heritability (h^2_{miss}).

2.1.4 Predicting the risk of disease using SNP data

Another common question explored in GWA studies is whether common genetic variants can be used to build prediction tools that may be relevant to clinical usage. This question becomes even more important in the case of rare diseases like MacTel. In fact, as already mentioned the first visual symptoms of MacTel are a direct consequence of structural changes in the retina, which often already involve photoreceptor death. Given the lack of treatment available for these structural changes, prevention and early diagnosis of the disease is key. Prediction is particularly important for prevention. If a prediction model adequately predicts the likelihood of a subject developing a disease, prevention measures can be adopted and recognition of early signs of the disease might be made easier. In our study, we were interested in predicting the likelihood of each subject being affected by the MacTel disease, based on their SNP information.

2.1.5 Finding candidate genes with expression quantitative trait loci

As mentioned in Chapter 1, on average only a third of SNPs associated with a disease act by influencing the expression of the nearest gene (cis-eQTLs). Other SNPs are more likely to affect the expression of more distant genes (trans eQTLs), or they affect the gene's final protein structure, instead of transcription levels, thus showing no association with transcript levels.

Exploring eQTL effects not only has the potential to shed light on the biological mechanisms underpinning the identified loci, but also allows prioritisation of specific genes of interest. We mined the GTEx database to further implicate the candidate genes and enzymes presented in **Figure 20**, but also to discover which were the potential gene effects by the identified loci and in which tissues these were most pronounced.

2.1.5 Exploring MacTel metabolic signal with metabolomics data

The results of the GWA study pointed to a possible genetic dysregulation of metabolomics signalling in MacTel patients. Specifically, we observed that 4 out of 5 SNPs were previously associated with abundances of circulating glycine and serine metabolites. As mentioned in Chapter 1, metabolomics is a growing field which has in recent years seen substantial methodological improvements in both machine measurements and bioinformatics techniques for the assessment of these measurements. A group of collaborators from the University College London and Moorfields Eye Hospital collected blood serum and plasma from unrelated MacTel cases and healthy controls. Both serum and plasma samples were then sent to

Metabolon Inc. to perform an ultrahigh-performance liquid chromatography-tandem mass spectroscopy (UPLC-MS/MS) which measured metabolomic abundances for hundreds of metabolites in each sample. The untargeted metabolomics dataset produced by the UPLC-MS/MS method was used to externally validate and explore the hypothesis of a disruption of glycine and serine metabolism in MacTel patients with respect to healthy controls.

2.2 Methods

2.2.1 Heritability

In 2011 Yang et al proposed a method implemented in the software GCTA to estimate the heritability of a trait (h^2) by using "all" SNPs (69). This method is based on the variance decomposition of a trait into a random term (all the SNPs) and a noise term. The methodology behind this approach is based on the construction of a genetic relationship matrix capturing similarities between all pairs of individuals.

If we assume that all causal variants for a trait are known, a genetic relationship matrix between individuals may be constructed using these causal variants and the fitted Linear Mixed Model (LMM). The method assumes that the phenotypic variability can be modelled using the formula:

$$var(y) = G\sigma^2_g + I\sigma^2_e$$

Where G is the genetic relationship matrix between individuals estimated using all causal variants. Since the causal variants are not known, GCTA uses, in the estimation process of G, all available genotyped SNPs. The matrix G is then multiplied by σ^2_g ,indicating the variance explained by the genetic contributions. The residual variance not explained by the genetics is modelled through an identity matrix I multiplied by the term σ^2_e, which indicates this residual variance. The complete methodology to estimate σ^2_g and σ^2_e can be found elsewhere (69). Given this formulation, GCTA defines heritability as

$$h^2 = \sigma^2_g / (\sigma^2_g + \sigma^2_e)$$

captures the proportion of the total phenotypic variance explained by genetic contributions.

This framework was developed to estimate h^2 only for quantitative traits. In our case, however, we have a dichotomous phenotype, where 0 and 1 indicate respectively the absence or the presence of MacTel. Work by Lee at al (72) presented a simple extension which allows heritability for dichotomous phenotypes to be estimated on the *liability scale*. This scale is constructed by fitting a *probit* distribution which assumes that the underlying trait risk is indeed continuous and normally distributed but can be dichotomized and transformed into the final observed phenotype (being affected with MacTel) by using a threshold. A classic example of a trait with an underlying *probit* distribution is fever, the presence of which is the result of thresholding over the continuous and normally distributed measure of human body temperature. However,

transforming a dichotomous variable on a liability scale requires the assumption of *disease prevalence* defined as the proportion of disease cases in the general population. Disease prevalence cannot usually be estimated using the proportion of cases versus controls in a GWAS setting as this would result in an extreme overestimation. MacTel prevalence is also difficult to estimate due to its rarity. In our case, we used two MacTel prevalence estimates: a prevalence of 0.004% as estimated by Klein at al. study (7) and 0.1%, as estimated by the Aung et al. study(8). We then explored heritability estimates using these two estimates which were treated as lower and upper bounds to MacTel prevalence.

2.2.2 Prediction of MacTel

A common method to predict disease status from genetics data is to use logistic regression and estimate the probability for each subject to develop a disease given their genotype. However, SNP chip data contains millions of variables (SNPs) that cannot be included simultaneously in classic regression models (73). A natural approach to this problem has been developed by Zou and Hastie in 2005 (73) called the Elastic Net method. It is a mixture of two previously described penalised regression models, the Lasso and Ridge regression models. The Elastic Net model tries to fit the data using the formula:

$$\hat{y} = X\hat{\beta} + \epsilon$$

Where X is a matrix containing all SNPs for all individuals and $\hat{\beta}$ is the vector of estimated regression coefficients. This vector is estimated by minimizing the following function:

$$\hat{\beta} = argmin_{\beta}(|y - X\beta|^2 + \lambda_1 |\beta| + \lambda_2 |\beta^2|)$$

In this function, λ_1 is the Ridge regression parameter and λ_2 is the Lasso regression parameter. Ridge regression "shrinks" the model parameters towards zero in order to avoid overfitting. On its own, ridge regression would include all the available variables (SNPs) and would be unable to fit the large number of SNPs which are usually much higher than the number of samples. The Lasso parameter takes care of this problem by shrinking or collapsing to zero the β parameters and thus selecting the "best" SNPs to predict a particular outcome. However, Lasso alone would fail to retain important variables that correlate with each other, as often happens in SNP chip data.

In order to perform this analysis, we used the R package GLMNET developed by Friedman et al. in 2009 (74) which reformulates the previous formula into the following:

$$\hat{\beta} = argmin_{\beta}(|y - X\beta|^2 + \lambda ((1-\alpha) |\beta|) + \alpha |\beta^2|)$$

λ is now a general penalization parameter and α becomes the parameter which balances the weight between the ridge regression ($\alpha = 0$) and the lasso regression ($\alpha = 1$). We tried different combinations of λ and α to predict MacTel status using the best combination of available SNP information. Parameter tuning was

performed by assessing prediction power using the Area Under the Curve value. To test that overfitting was not occurring we split the discovery dataset into a training dataset (75% of observations) and a validation dataset (25% of observations) and assessed whether similar prediction power eventuated, even though parameter tuning was performed separately in both the training and the discovery set both.

2.2.3 eQTL analysis

The general framework of eQTL studies has been presented in Chapter 1. In eQTL settings, the association between a SNP of interest and the expression of a gene is usually tested through a linear regression

$$\widehat{e_g} = SNP_s \widehat{\beta}_{sg} + \epsilon$$

$e_g$is the expression of gene g and $\hat{\beta}_{sg}$is the estimated additive effect of each copy of the minor allele of SNP s. A SNP is deemed to significantly affect the expression of a gene if the p-value related to the rejection of the null hypothesis $H_0 : \hat{\beta}_{sg} = 0$ falls below a certain threshold (typically much lower than 0.05, to correct for the many tests that are performed in eQTL studies). However, both gene expression and SNP genotype calls are often affected by noise from different sources, such as for example SNP chips, population stratification, RNA sequencing platform, sample processing batches. To test which genes' expression levels were affected by the SNPs identified as affecting MacTel risk, we mined the GTEx database (75) which take this technical noise into account by pre-processing both SNP genotypes

and gene expression matrices for each tissue. Additionally, GTEx tests the association between SNPs and genes by including several covariates into the regression model. The specific methodology used by GTEx database can be found elsewhere (51).

In mining the GTEx database to find which genes were affected by the MacTel associated SNPs, we first selected all SNPs from the GWAS with association p-value falling below the 1e-5 threshold (suggestive significance). To ensure we would find all the associations for each locus in the GTEx database we searched for all SNPs in high LD ($R^2 > 0.8$) with the selected SNPs that might be used as a proxy. To this end, we uploaded the list of rs identifiers of SNPs on the online SNAP tool (76). Using all the proxy SNPs as well as the original SNPs of interest we mined the GTEx v6.0 database to identify any significant cis-effects of these SNPs in all tissues available on the database (44 tissues). eQTL significance varied from tissue to tissue due to the differences in sample sizes. More information about tissues sample sizes as well as significance assessment in the GTEx database can be found elsewhere (51, 75). For ease of interpretability, GTEx eQTL effects sizes were flipped if the SNPs allele used by GTEx was the opposite of the MacTel risk allele.

2.2.4 Metabolomics analysis

We analysed two untargeted metabolomics datasets (generated by our collaborators Dr Marcus Fruttiger, UCL, and Dr Catherine Egan, Moorfields' Eye

Hospital), one derived from blood serum, with the other from blood plasma. The specific methodology used to process the blood sample and measure the metabolomics abundances can be found later, in Chapter 4. . The serum-derived dataset contained 50 healthy individuals and 50 MacTel patients, with all subjects reported to be genetically unrelated to each other. Metabolic abundances for 1,281 metabolites were measured for each individual by Metabolon Inc.. The missingness rate among cases and controls was calculated for each metabolite. Although Metabolon Inc. provided imputed and normalised metabolomics abundances, metabolites with an original missingness rate higher than 10% in either cases or controls were excluded. Since missing values are indicative of an extremely low concentration of a metabolite in a specific sample, metabolomics imputation was performed by imputing the minimum of the observed abundances across all samples for each metabolite. Metabolomics abundances were assessed for variation across samples to ensure no redundant metabolite was used for the analysis. To this end, metabolites with an extreme density concentration around the mode were discarded. To further limit multiple testing, we discarded all metabolites which a correlation coefficient with at least one other metabolite, greater than 0.95, since these were deemed to convey redundant biological information. Outlier subjects were identified by estimating the first two principal components with the remaining metabolite data and visually identifying outlier individuals on the principal components' projection. Metabolomics abundances were further normalised across samples using quantile normalisation as defined by the package limma (77) in R. Since cases and controls were well balanced for different factors such as sex at birth, diabetes status, age, and ethnicity,

differential abundance analysis was performed with a univariate two sample t-test assuming equal variances. Inflation of T-test p-values was controlled for using a 'genomic inflation'-like procedure. In this particular publication, the procedure involved the estimation of an inflation coefficient by regressing the observed p-value distribution against a theoretical one by excluding the top most significant 10% metabolites. Each p-value was then divided by the inflation coefficient. Glycine and serine were then assessed for significance. Significance despite multiple testing was further assessed by comparing the observed p-values of such metabolites against the inflation-controlled Quantile-Quantile (QQ) plot of the entire metabolomics data.

2.3 Results

2.3.1 MacTel Heritability

2.3.1.1 MacTel Total heritability

The estimated heritability using all directed genotyped SNPs after QC filtering (~1M) and assuming a disease prevalence of 0.1% was h^2 =0.742. Hence 74% of the MacTel phenotypic variability on the liability scale could be explained by the genetic influence of SNPs. Interestingly, when assuming a prevalence of 0.0045% the estimated heritability collapsed to h^2 =0.215, indicating that only 21% if the phenotypic variability could be explained by genetic contributions. Given the extreme differences between these two estimates, we empirically explored how MacTel estimated heritability changed according to prevalence rate. The result is

presented in **Figure 21**. This analysis clearly demonstrated how the heritability is affected if the population prevalence fell below the 0.02% first assumed. We additionally tried to estimate the total heritability by using all imputed SNPs (~6M). Interestingly, by assuming a prevalence of 0.1% the estimated total heritability was reduced to h^2 =0.412.

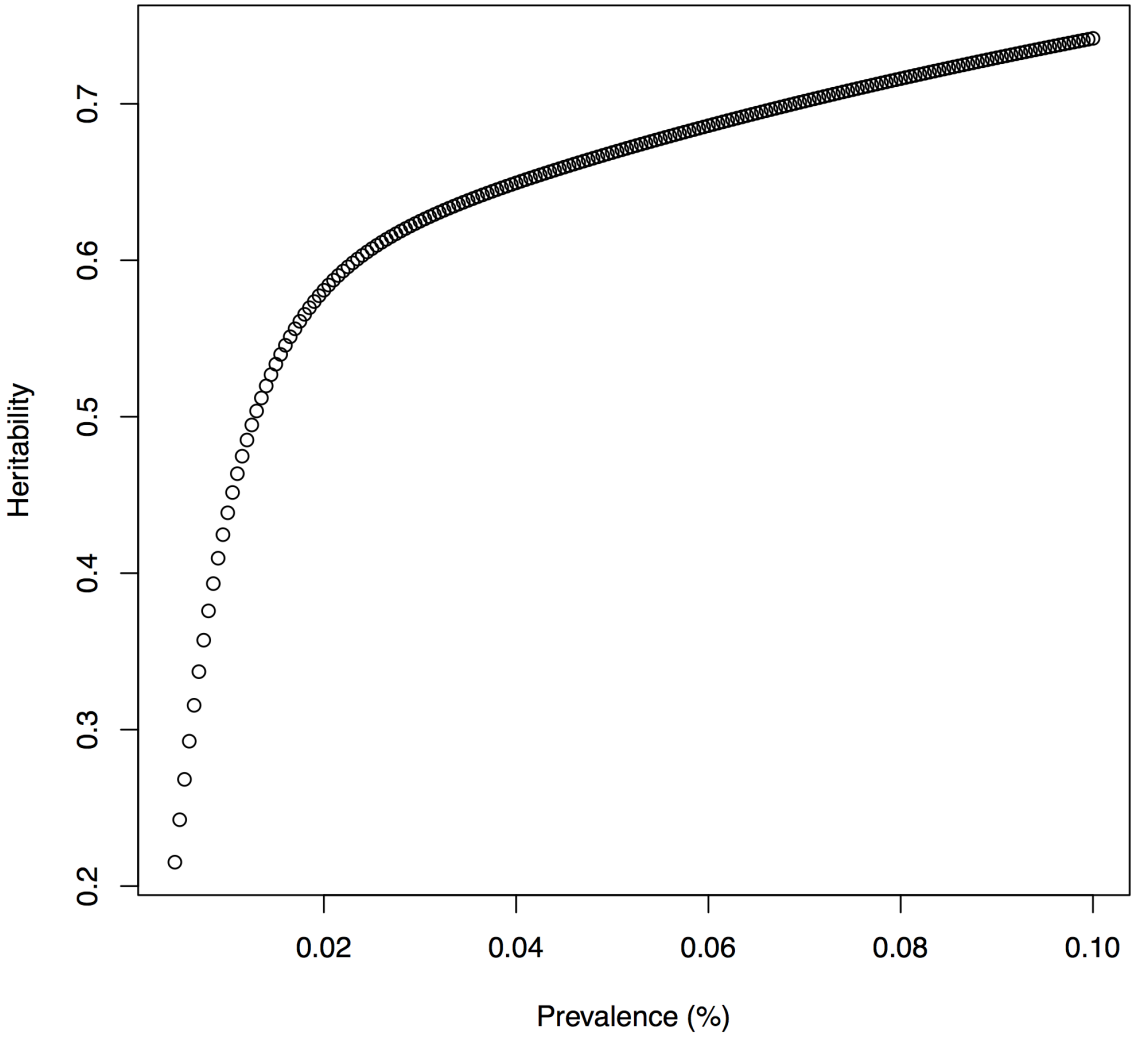


Figure 21: Total narrow-sense heritability estimated by assuming different disease prevalence ranging from 0.0045% to 0.1%.

2.3.1.2 MacTel known, explained, and missing heritability

To estimate the known heritability we used the 5 SNPs tagging the replicated loci that were identified by the GWAS analysis rs73171800, rs715, rs477992, rs9820286, and rs4948102 tagging loci 5q14.3, 2q34, 1p12, 3q21.3, and 7p11.2 respectively. The known heritability explained by these loci assuming a prevalence of 0.1% was h_{known}^{2} =0.039. Unsurprisingly, assuming a heritability of 0.0045% resulted in a drop of estimated known heritability to h_{known}^{2} =0.011. These estimates were then used to calculate the proportion of heritability explained by known loci, by dividing h_{known}^{2} by h^2. By using both disease prevalences we obtained an explained heritability of $h_{explained}^{2} = 0.053$ translating to a missing heritability of $h_{missing}^{2} = 0.947$. This meant that only 5.3% of the total genetic heritability could be explained by the 5 replicated SNPs leaving unexplained or "missing" the other 94.7%.

2.3.2 Predicting MacTel using SNP data

A simple prediction tool constructed using logistic regression and trained on the 5 replicated SNPs resulted in an Area Under the receiving operating Curve (AUC) of 0.71. The same model lost some prediction power on the replication dataset used to replicate such SNPs resulting in an AUC of 0.68. Although the prediction power of this model was relatively large compared to other prediction tools using genome-wide significant SNPs on different polygenic diseases, we tried to improve the

prediction power by employing Elastic Net models. The replication dataset of 172 cases and 1,134 controls only contained the 5 SNPs that needed to be replicated. Since the elastic net would select many more than these five SNPs, we divided the discovery dataset of 476 MacTel patients and 1733 controls further into a training dataset containing 75% of the original discovery data and a validation set containing the other 25% of the original dataset. For each combination of α and λ we assessed the prediction power as AUC on a 20-fold cross-validation of the training data. For each value of α we selected only the topmost predictive values of λ. The AUC results from the different models on the training data are presented in **Figure 22**.

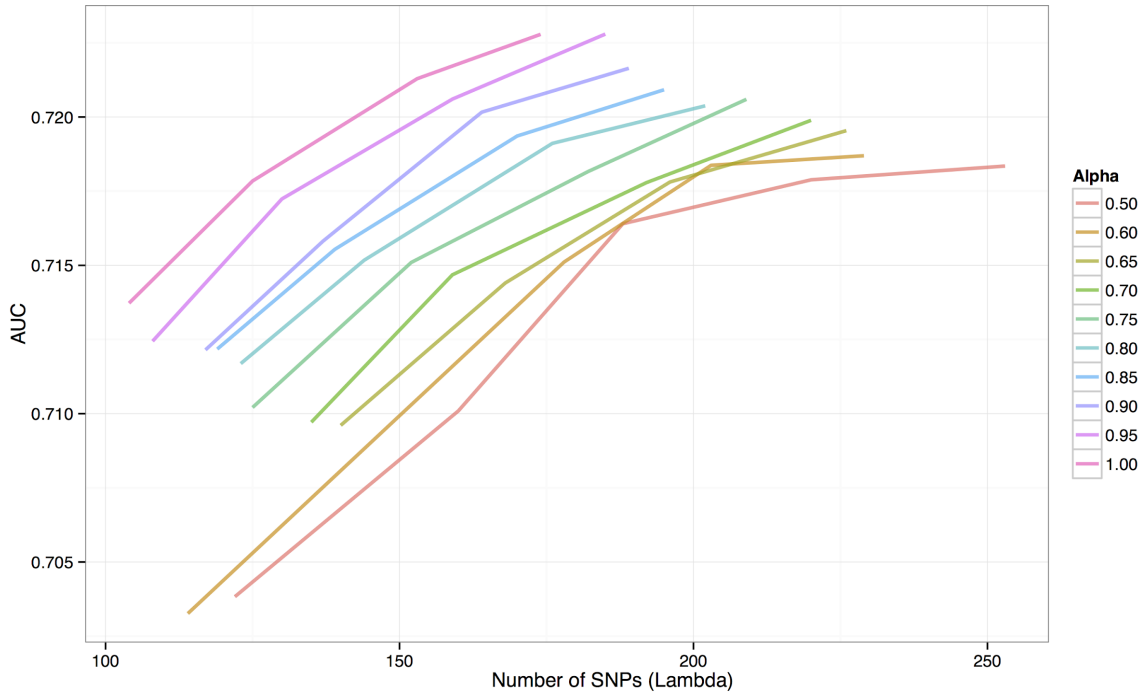


Figure 22: Prediction power as described by AUC value of different elastic net models on training data.

The best performing prediction model, whilst keeping the minimum number of SNPs was a complete Lasso model ($\alpha = 1$) with $\lambda = 0.0325$. This model used 174 SNPs and reached an AUC of 0.723. When assessing the prediction performance of the same model on the whole training dataset the model reached an AUC of 0.972. However, when assessed on the validation set the prediction power dropped to 0.677. The receiving operating curves under which these AUCs were calculated are shown in **Figure 23**.

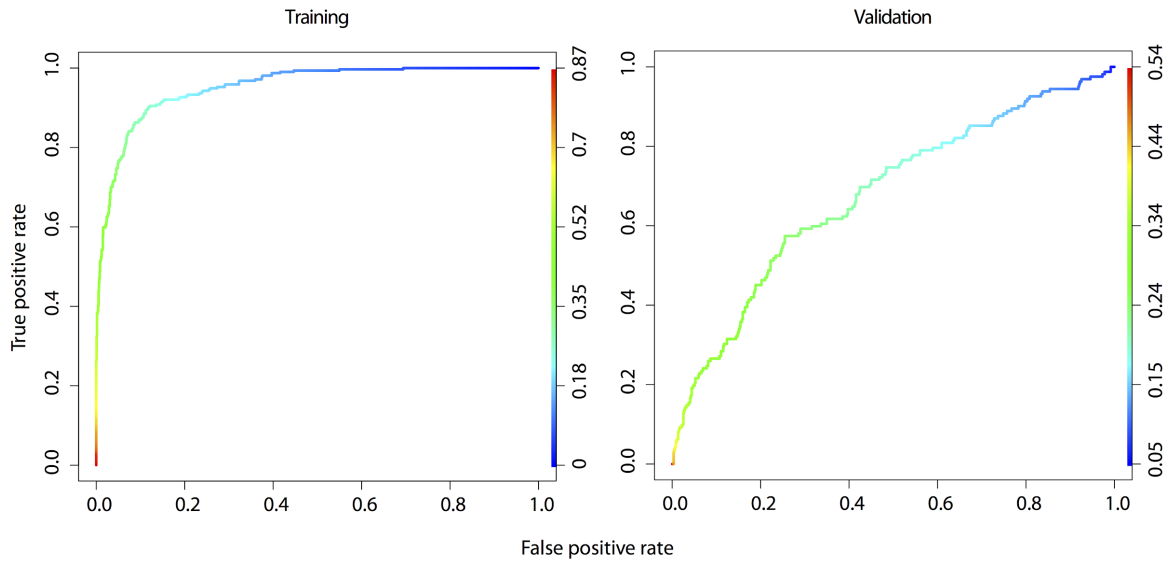


Figure 23: Prediction power comparison of selected lasso model between training and validation set.

Lower data folding (5-folds and 10-folds) on the training dataset were additionally tried to prevent overfitting. However, none of these models resulted in better performance on the validation set. We additionally tried to train a similar model by also including all SNP-SNP interactions between the SNPs selected by this model as well as forcing the model to always select the 5 replicated SNPs to be among the chosen SNPs. However, these attempts also resulted in a very high

prediction power on the training set and extremely low power on the validation set (results not shown).

2.3.3 Candidate genes identified from mining the GTEx eQTL database

Thresholding the SNPs associated with the disease with a nominal p-value less than 1e-5 resulted in 597 SNPs tagging 24 different loci. The SNAP tool did not return any new SNPs in LD (R^2>0.8) with the initial 597 SNPs. When mining the GTEx database v6.0 we identified 240 of these SNPs to have significant cis-eQTLs, with expression in at least one tissue for each gene. These 240 SNPs were tagging 11 out of the 24 loci initially identified. A visual representation of the significant eQTLs found is presented in **Figure 24**.

Figure 24: Significant eQTL results of the genome-wide significant and suggestive loci in the GTEx database v6.

Among the 5 replicated SNPs of interest, we found MacTel risk SNPs in locus 1p12 (chr1: 120208Kbp - 120281Kbp) to affect two genes: *ZNF697* and *PHGDH*. Interestingly, the latter of these was affected in a positive direction in the Skin and in a negative direction in Testis. SNPs in locus 5q14.3 in chromosome 5, presenting with the biggest effect in the GWAS analysis, were found to significantly increase the expression of the gene *TMEM161B-AS1* in 20 tissues. Locus 7p11.2 in chromosome 7 contained SNPs affecting transcription of *CCT6A*, *GBAS, PSPH*, *NUPR1L* and *SUMF2* in different directions, depending on the gene. Lastly, two significant eQTL loci were identified for the loci 2q34, on chromosome 2, and 3q21.3 on chromosome 3. SNPs in the other six suggestive loci were found to affect the expression of the genes *SH3YL1*, *ACP1*, *RP11-222K16.1, AC098973.2, LEPREL1, TTC39B, SLC31A1, NRBF2, REEP3, TRAV19, RP11-49G10.3*, and *ACTL10*.

2.3.4 Metabolomics results

The metabolomics sample comprised of 50 MacTel cases and 50 healthy controls was well matched for sex (25 males and 25 females in both groups), age (cases were on average 64 years old while controls were 63 years old on average), type 2 diabetes status (12 diabetics patients and 38 non-diabetics in both groups), and ethnicity (47 Caucasian controls, 48 Caucasian cases). Of the initial 1,281 metabolites, 403 were excluded due to high missingness in either control or case groups, 49 were excluded due to extremely high abundance density (or low dynamic range) over the mode, and 30 were excluded due to high correlation with

at least another metabolite. The remaining 799 metabolites were used to assess sample quality. A visual representation of all subjects with the first two principal components calculated from all remaining metabolites is presented in **Figure 25**. No outliers were identified.

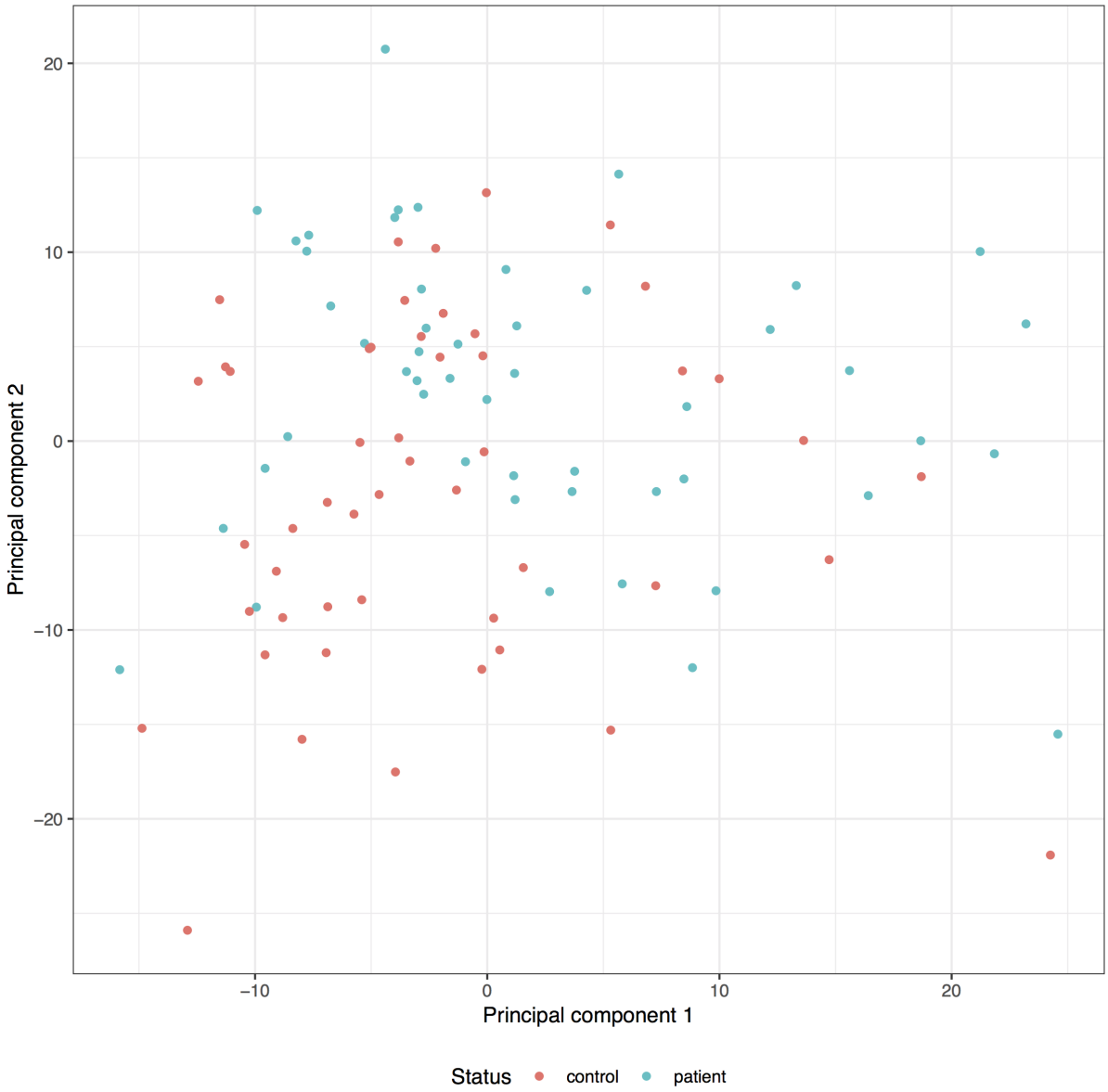


Figure 25: Samples position on the first two principal components projection of metabolomics data. No identifiable outliers.

Differential abundance testing revealed glycine (p=3.5e-10, T= –6.98), and serine (p=4.8e-7, T= –5.39) to be the first and third most differentially abundant metabolites with both achieving multiple testing adjusted significance. The second most significant metabolite was threonine (p=1.6e-7, T= –5.64). By regressing the observed p-value distribution onto the theoretical quantile of a uniform distribution we calculate a p-value inflation factor of 1.73 for all tests (displayed by the red line in **Figure 26**).

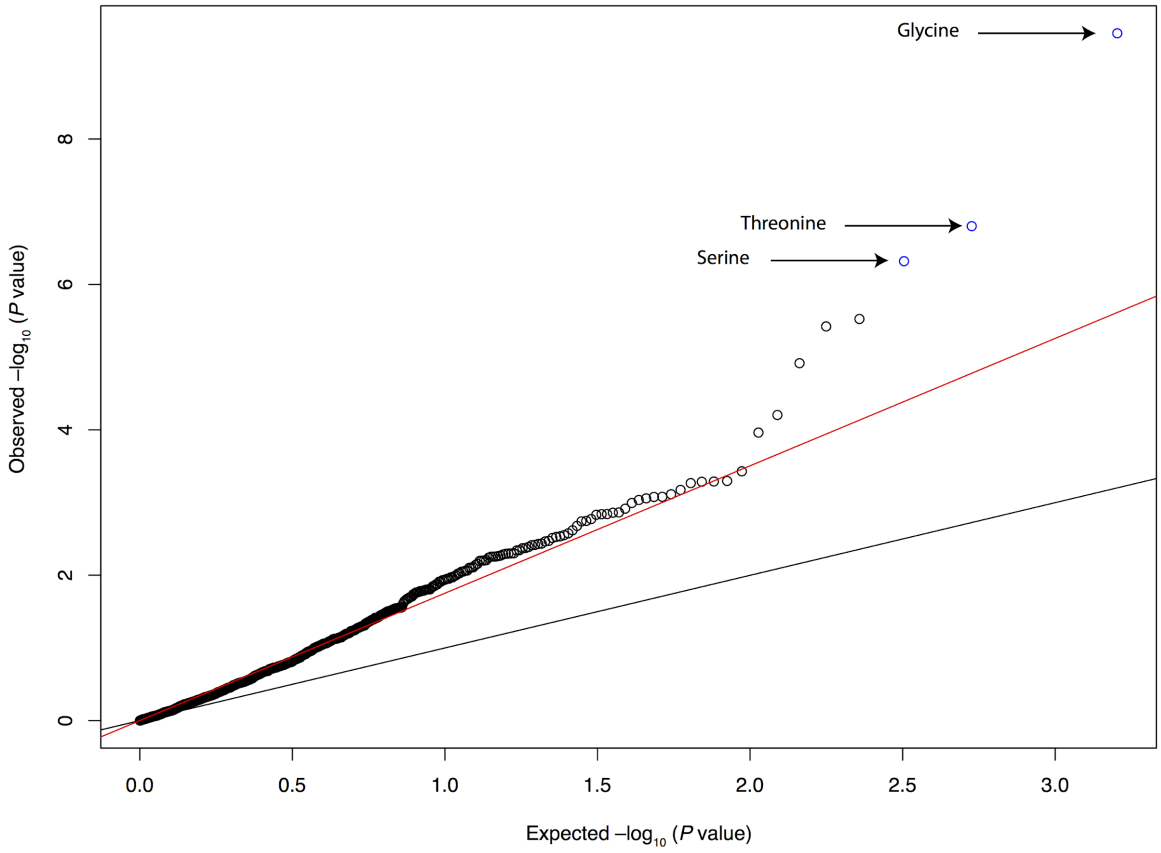


Figure 26: Quantile-quantile plot of the observed and expected p-value form metabolomics analysis corrected by genomic inflation like coefficient.

By correcting the observed p-values for the inflation factor we obtain p=4.0e-6 for glycine, p=1.3e-4 for threonine, and p=2.5e-4 for serine. The deviation of these three metabolites from the genomic inflation corrected identity line highlighted how the significance of these three metabolites would be retained even after correcting for the multiple testing burden and correction for inflated -log10(p-values). The same metabolites were also found to be significant, or borderline significant, if we corrected the inflation corrected p-values for multiple testing using the Benjamini-Hochberg procedure. We observed all three metabolites to be depleted in MacTel patients compared to controls with a log2FC of -0.54, -0.35 and -0.38 respectively for glycine, threonine and serine. The distribution comparison between cases and controls of the log2 normalized metabolomics abundances for these metabolites is shown in **Figure 27**.

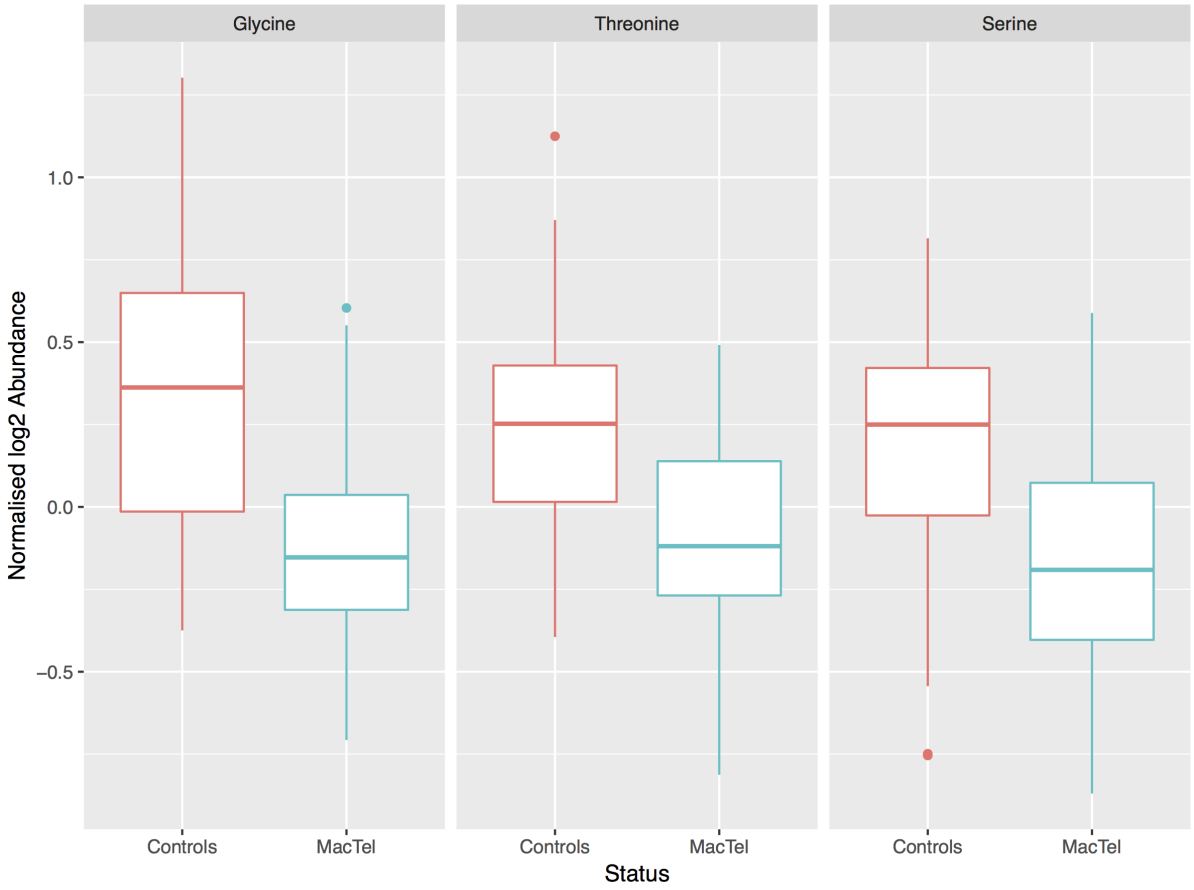


Figure 27: Normalised log2 abundance of glycine, serine and threonine grouped by MacTel case and control status.

2.4 Discussion

2.4.1 MacTel appears to have a substantial genetic contribution to its phenotypic variability, driven by additive polygenic contributions

The GWAS results identified MacTel as a polygenic rather than a monogenic Mendelian disease for the first time. The heritability analysis estimated the contribution of the genetic polygenicity that might be contained by common

genetic variants explain the largest portion disease risk variance. We found that MacTel heritability could not be estimated precisely, with our estimates ranging from 0.21 to 0.74. This large range was a result of the very different prevalence estimates used. However, as shown in the sensitivity analysis in **Figure 21**, most of the heritability estimates were contained between 0.6 and 0.7, indicating that, although the difference in heritability estimates was very large, the lowest appears to be more an extreme lower bound estimate. This large difference in heritability may arise from the fact that the liability scale is assumed to follow a normal distribution of disease risk (here modelled as prevalence). Such an extremely low prevalence would, in fact, sit on the tail of the normal distribution resulting in an extremely low heritability estimate. In general, MacTel heritability results show strong support for the hypothesis that MacTel can be largely explained by genetic influences.

When including all the imputed SNPs surviving GWAS QC steps we observed a drop in the estimated heritability. This problem has been addressed by Chen et al (68) who noticed that the GCTA framework is not robust with the inclusion of non-causal variants, stating in their simulation study that "the addition of non-causal SNPs seems to have partly swamped the LMM's dissection of signal from noise". We also recognised that this heritability estimate could be further increased by accounting for the fact that MacTel causal variants may have a MAF smaller than the retained SNPs, as discussed by Yang et al (69). However, the small sample size available precluded us from performing analyses that would allow the identification of low MAF variants associated with the disease.

The proportion of heritability explained by the five replicated loci represented only 5% of the total heritability. This is consistent with previous work which indicates that most of the causal SNPs (or the genotyped SNPs in LD with the causal ones) fall below the genome-wide significance threshold and are often missed by GWA studies. It has been suggested that an infeasible amount of data would be required to discover all the causal SNPs for a polygenic trait such as MacTel disease, in order to fully account for the missing or unexplained heritability (68).

We also recognised that in this thesis we have not addressed the problem of "phantom heritability" and "heritability estimated by all pseudo-significant SNPs" which tries to estimate the heritability with all the SNPs that achieve a p-value less than 0.05. These analyses were not performed since they were not deemed necessary for the Scerri at al publication.

To conclude, we found that MacTel has a substantial genetic component that explains a large percentage of disease heterogeneity. However, the candidate loci individuated so far by the GWA study only account for a small fraction of this heterogeneity and more powerful GWAS are needed to identify additional loci for the disease.

2.4.2 Prediction of MacTel disease based on SNP chip data is promising, but not yet clinically relevant

With the heritability analysis revealing a strong genetic component affecting disease heterogeneity, we sought to assess the predictive power of using the 5 replicated SNPs tagging the loci of interest. Using only these 5 SNPs, we were able to reach an AUC of 0.7 in both training and external validation set. When we tried to include more SNPs, using an Elastic Net method, this resulted in greater prediction power on the training dataset, but no improvement in the validation dataset. This highlighted a problem of overfitting of our model in the training data, most likely caused by the relatively small sample size of the discovery GWAS data, which was additionally reduced in order to separate training from the validation set. We acknowledge that there are advanced methodologies implemented in recent years to address this problem. However, the genetic risk prediction of MacTel was not the main goal of our study but rather more of an exploratory analysis, therefore we decided not to pursue any additional and more complex analyses.

To conclude, we found that common genetic variants were able to estimate a reliable risk score for MacTel disease. However, as already highlighted by the heritability analysis, more studies with bigger sample sizes will be needed to improve the prediction tools for this disease based on SNPs.

2.4.3 eQTL analysis reveals effects on genes known to affect metabolite abundance and eye disease genes

When mining the GTEx V6.0 database for candidate genes involved in MacTel we found loci 1p12 and 7p11.2 to respectively affect the expression of genes *PHGDH* and *PSPH*, which, as shown in **Figure 21**, are both ubiquitously linked to glycine and serine metabolic biosynthesis. Interestingly, we could not find any significant eQTL effect of the SNPs in locus 2q34 on gene *CPS1* despite these SNPs being located on the 3' UTR region of this gene. However, there are many ways in which SNPs may affect the functional properties of a gene apart from influencing its transcription. For example, SNPs often tag multiple nucleotide changes as they usually belong to long haplotypes. If many nucleotide changes are happening in a specific haplotype, the rate at which a gene is transcribed might remain invariant but the final 3D structure of its protein might be significantly different. Correct 3D structure of proteins translates to the proper functioning of such proteins. Through this mechanism, SNPs might affect the biological functioning of a gene's protein product via its structure, rather than the gene transcriptional magnitude. Hence, it might be speculated that the SNPs identified in locus 2q34 might be responsible for structural changes in the CPS enzyme, catalysed by *CPS1*, impairing its functioning and resulting in an alteration of the metabolic biosynthesis of glycine and serine. Lastly, we also found significant eQTL effects for the identified SNPs in locus 5q14.3 on gene *TMEM161B-AS1* which is a gene whose function is currently unknown.

Among the suggestive significant SNPs, we found some other genes of interest. SNPs in locus 3q28 were found to have eQTLs with *LEPREL1*. This gene, also known as *P3H2*, is a member of the prolyl 3-hydroxylase subfamily of 2-oxoglutarate-dependent dioxygenases and has been observed to be associated with non-syndromic severe myopia with cataract and vitreoretinal degeneration. Another interesting gene found to have eQTL in locus 9p22.3 was *TTC39B* which has been observed to regulate HDL cholesterol metabolism. Lastly, gene *SLC31A1* which we observed to have an eQTL in locus 9q32 has been linked to Wilson's disease, a copper related pathology which often results in cellular damage of the eye.

By mining the GTEx database, we were able to identify different genes whose expression might be modulated by the SNPs identified in our GWAS analysis. Interestingly, some of these genes were related to serine and glycine biosynthesis, whilst others were related ocular phenotypes. However, it is important to realise that the GTEx database V6.0 did not contain any eQTL data for retina and that a specific retinal tissue analysis to confirm such results would be required to validate the associations observed in this study.

2.4.4 Semi-targeted metabolomics analysis reveals disruption of glycine, serine and threonine metabolic pathway

The emerging hypothesis from the GWAS results highlighted a possible metabolomic disturbance in the glycine and serine biological pathway in MacTel patients. By performing an exploratory analysis in the untargeted metabolomics data, we observed glycine, serine and threonine to be the top 3 most significant metabolites to distinguish between cases and controls. This result supported the emerging hypothesis of a metabolomics aberration - arising from genetic disturbances - on MacTel risk. Although the SNPs identified in the GWAS analysis only suggested glycine and serine as metabolites whose abundance might be affected in MacTel patients, we discovered that threonine was also similarly depleted. Moreover, we observed that glycine, serine and threonine were all part of the same biological pathway in mammals as displayed by the KEGG pathway shown in **Figure 28**.

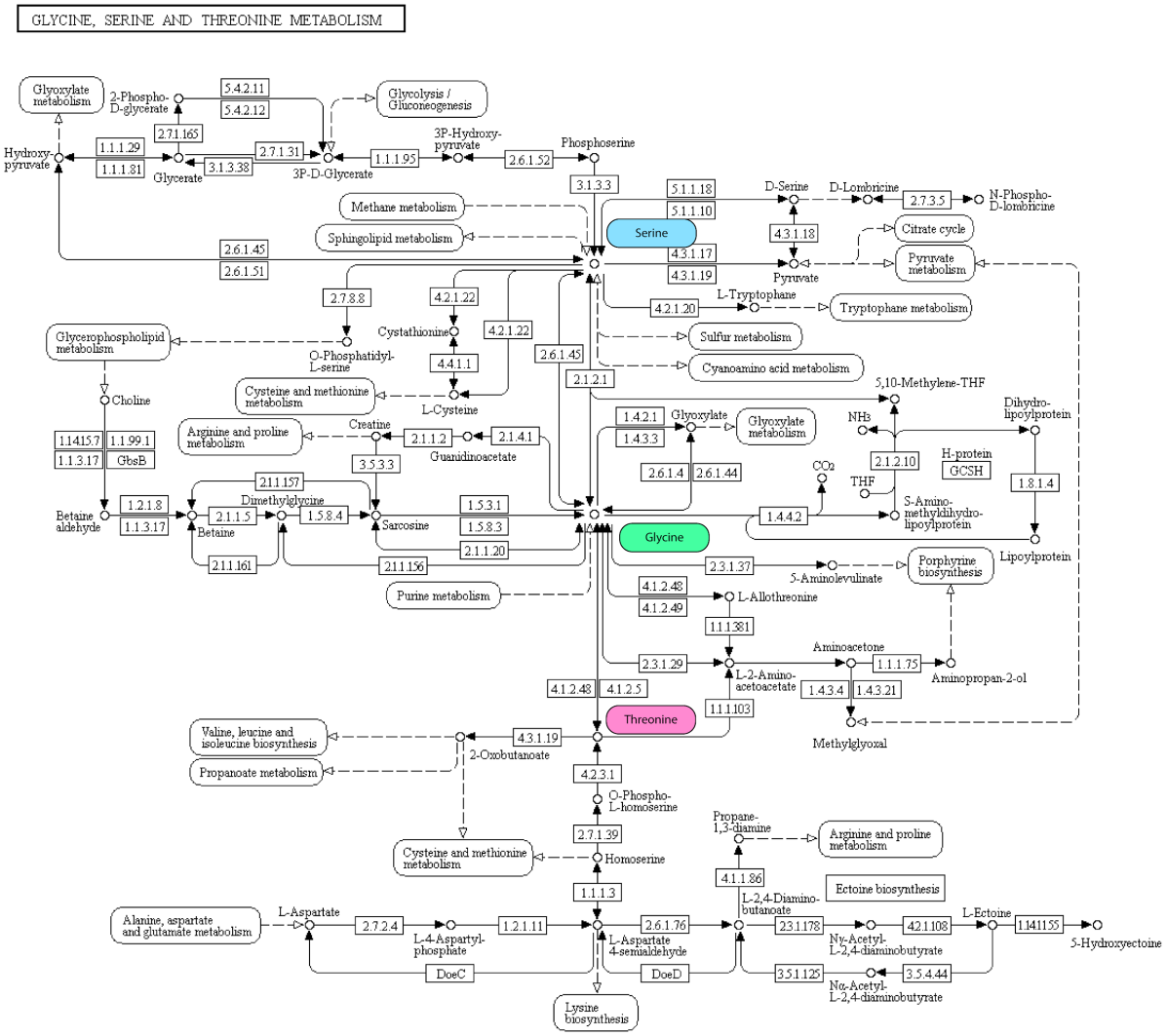


Figure 28: KEGG pathway for "Glycine, Serine and Threonine Metabolism" visualising metabolic connections between glycine, serine and threonine.

Given the translational relationship between glycine and threonine, as well as glycine and serine, this result suggests that depletion in threonine might be a consequence, or bystander effect, of the genetic perturbation of glycine and serine among MacTel patients identified by our study.

This analysis only focused on the validation of the main metabolites of interest and did not examine the other metabolomics dysregulation that might have been

identified with an in-depth untargeted metabolomics analysis. Such an analysis is presented in **Chapter 4**.

2.5 Conclusion

Our study was the first GWAS ever performed for MacTel. A schematic of the study findings is presented in **Figure 29**. This figure will be updated at the end of each research chapter to summarise the main findings and disease drivers highlighted by this thesis. The GWAS analysis resulted in five genetic loci, four previously linked to glycine and serine which were in fact observed to be depleted in MacTel patients through a metabolomics study. The analysis also revealed that MacTel is a highly heritable trait and also indicated that the identified SNPs had an elevated prediction power although not high enough to be of clinical usage. Among the identified SNPs we also confirm the presence of eQTL on genes implicated in glycine and serine biosynthesis as well as on other pathways whose dysregulation might result in further ocular damage.

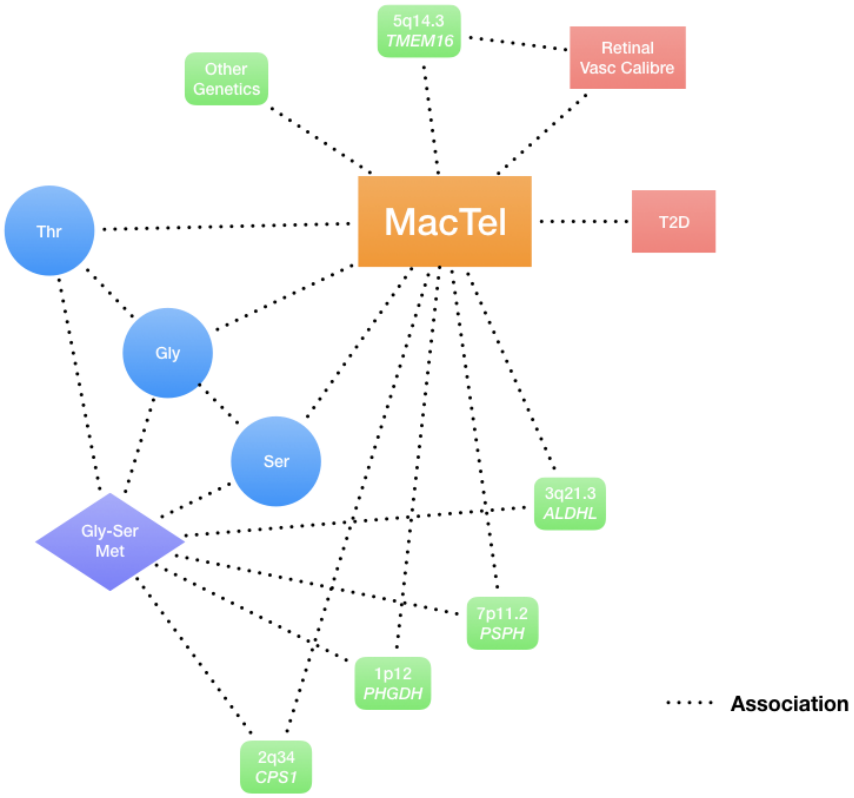


Figure 29: Main findings schematics displaying the drivers and traits associated with MacTel. Genetic traits are displayed as green rounded squares. Metabolites are displayed as blue circles. Metabolic pathways are displayed as purple diamonds. Lastly, remaining phenotypic traits are displayed as red squares. Found associations are represented by dotted black lines.

3 Dissecting MacTel genetic signals to understand disease heterogeneity and aetiology

3.1 Introduction and study aims

The results presented in chapter 2 clearly highlighted the need for a deep investigation of specific genetic components that might affect metabolic biosynthesis or retinal vasculature traits, and through these, impact the disease insurgence. Moreover, the advancement of bioinformatics tools and resources allowed us to explore more deeply this complex disease which the GWAS dataset only scratched the surface of.

The following chapter will firstly introduce specific concepts related to MacTel heterogeneity arising from different sources like genetics, metabolomics and phenomics. It will then present a publication describing an in-depth analysis of MacTel genetic signatures and how we integrated that information with publicly available datasets as well as with phenomics retinal data acquired on MacTel patients to discover a possible causal mechanism for the disease aetiology and its progression as well as the prioritization of new genetic loci of interest. It will then discuss the findings of this research and connect it with the next chapter.

3.1.1 MacTel is a complex disease

Our initial GWAS study presented the first clue of a genomic and metabolomic component on MacTel disease. However, one of the findings of that research was that MacTel is likely to be a genetically complex disease. Specifically, we found that some loci seemed to be involved with the biosynthesis of glycine and serine while other loci seemed to be involved with retinal vascular calibre. However, our previous results on MacTel heritability highlighted that this complexity should be contained and MacTel variability can be mostly explainable by common variants.

The same study also highlighted how MacTel patients presented with alterations in metabolic abundances of serine, glycine and threonine. However, other metabolites might be important for MacTel due to connections with glycine and serine pathway or because of different mechanism predisposing for the disease.
Lastly, as presented in Chapter 1, MacTel is characterised by a multitude of clinical signs detectable through different imaging techniques. Although some of these are fundamental for MacTel diagnosis and form part of the core of the retinal phenotypes presented by each patient, some of these signs are unique and might differ from patient to patient.
This finding encourages a deep investigation on how the heterogeneity observed in different 'omics data might be connected with each other and how analyses

performed to uncover connections between these 'omics layers might inform on disease mechanism.

3.1.2 Missing data imputation concept

Phenomics information on MacTel patients used in this study came from the Natural History Study and Registry of Macular Telangiectasia (NOHS) (10). This data contained longitudinal retinal phenotypic observations (colour fundus images, optical coherence tomography and fluorescein angiography images) that were collected over the period 2005-2015 from 1,716 patients (3,410 eyes in total). This very large dataset presented, like many other longitudinal datasets, a large amount of missing information. The missingness arose mainly to either imperfection of the retinal images or was related to changes in retinal photography machines. The problem of missing data and how to handle such missingness has been extensively explored in epidemiology (78) and a review of such a problem would be beyond the scope of this thesis. For this study, we decided to impute the missing data. Again, much research has been performed on the topic of missing data imputation (79) with most methodology trying to address missingness arising from different scenarios. Missingness has been described to arise from three different scenarios (79):

1. Missingness completely at random: this missingness arises randomly and does not depend on anything related to the observations (e.g. machine randomly not working)

2. Missingness at random: this missingness can arise from any factor apart from the ones which are missing (e.g. machine working badly on children)
3. Missingness not at random: this missingness arises depending on what value would have been measured (e.g. the machine cannot detect values smaller than a threshold)

For our study, we assumed that missingness was "at random", implying that the missing did not depend on the actual value of the missing entry. Modern imputation methods often rely on the observed relationship between the variable containing missing values and other variables. This relationship can be used in a prediction setting and imputation can be performed by predicting the missing data using all other available information. However, further complications arise when missing values are represented in multiple variables which might be used to impute each other. A well-known method to overcome such a complication is known as the Multivariate Imputation by Chained Equations (MICE) (80). This is a well-known method which was cited more than three-thousand times at the time of this thesis.

3.1.3 Constructing endophenotypes using factorial analysis

The NOHS contained a very large number of phenotypes that would have also posed a dimensionality problem. To reduce the number of variables, we hypothesise that some of the observed phenotypes are different observations of

common endophenotypes. The definition of endophenotype is not uniform across the literature (81). However, endophenotypes have been defined as “measurable components unseen by the unaided eye along the pathway between disease and distal genotype” (82). Other definitions of endophenotypes comprise the notions that endophenotypes should be heritable, quantitative, associated with the disease, cosegregate in families where a disease is observed and may be related to causative mechanism (81). These definitions support the idea that the MacTel clinical signs, here called phenotypes, may be consequences of common endophenotypes. If a group of phenotypes is produced by the same endophenotype, it is assumed that these phenotypes are also correlated with each other. Based on these assumptions we decided to create these endophenotypes from the initial phenotypes using a technique called factorial analysis. Factorial analysis examines all the correlations between the phenotypes and, based on these, tries to discover and create endophenotypes by “melting” together the initial phenotypes. If correctly performed, this process is able to substantially reduce the dimension of a dataset by, at the same time, creating interpretable endophenotypes that “summarise” the observed phenotypes.

Below is a preprint of my paper that describes this, and additional analyses.

3.2 Extract from Bonelli et al, Genetic Disruption of Serine Biosynthesis is a Key Driver of Macular Telangiectasia Type 2 Aetiology and Progression

Note: The following manuscript contains figures which are indexed differently from the rest of the thesis as this subchapter is presented in the same formatting as the submitted manuscript.

Genetic Disruption of Serine Biosynthesis is a Key Driver of Macular Telangiectasia Type 2 Aetiology and Progression

Roberto Bonelli,[1,2] Brendan R E Ansell,[1,2] Luca Lotta,[5] Thomas Scerri,[1,2] Traci E Clemons,[3] Irene Leung,[4] The MacTel Consortium,[6] Tunde Peto,[7] Alan C Bird,[8] Ferenc Sallo,[9] Claudia Langenberg,[5] and Melanie Bahlo*[1,2].

1 Department of Medical Biology, The University of Melbourne, 3052, Parkville, Victoria, Australia.
2 Population Health and Immunity Division, Walter + Eliza Hall Institute of Medical Research, 3052, Parkville, Victoria, Australia. 3 The EMMES Corporation, Rockville, 20850, Maryland, United States.
4 Department of Research and Development, Moorfields Eye Hospital NHS Foundation Trust, EC1V 2PD, London, United Kingdom.
5 MRC Epidemiology Unit, University of Cambridge, CB2 0SL, Cambridge, UK.
6 A list of members and affiliations appears in Table S7.
7 Department of Ophthalmology, Queen's University, Belfast, BT7 1NN, United Kingdom.
8 Inherited Eye Disease, Moorfields Eye Hospital NHS Foundation Trust, EC1V 2PD, London, United Kingdom.
9 UCL Institute of Ophthalmology, EC1V 2PD, London, United Kingdom.
* Correspondence: bahlo@wehi.edu.au, +61 03 9345 2555

Abstract

Macular telangiectasia type 2 (MacTel) is a rare heritable degenerative retinal disease, often comorbid with type 2 diabetes, that causes blindness and is largely untreatable. We previously found genetic loci associated with MacTel linked to retinal vasculature calibre and the glycine, serine and threonine metabolic pathway.

We piece together causal genetic contributions to MacTel and its consistent retinal phenotypes, by integrating disease genetic markers with those of vascular and metabolic traits, and apply advanced statistical genetics techniques including Mendelian randomization, MTAG, conditional and interaction genome-wide association analysis, and genotype-phenotype analyses.

We identify serine to be a main causal driver of disease incidence and progression, with a lesser, but significant, causal contribution of type 2 diabetes genetic risk. We further show that glycine, threonine and retinal vascular traits are instead unlikely to be causal for this disease. Through conditional regression analysis, we resolve three novel disease loci independent of endogenous serine biosynthetic capacity. We lastly demonstrate how disease loci cluster into functional groups affecting retinal endophenotypes.

Follow up studies after GWAS integrating publicly available data with deep phenotyping are still rare. Here we describe such analysis, where we integrate retinal imaging data with MacTel and other traits genomics data, performing an integrated analysis informing on the biochemical mechanisms likely causing this disorder. Our findings will aid in early diagnosis and accurate prognosis, as well as improving prospects for designing effective interventions and will serve as a useful template for such studies in other disorders post-GWAS.

Retinal Disease | Mendelian Randomization | Metabolomics | GWAS | Serine

Introduction

MacTel Disease and previous GWAS study

Macular telangiectasia type 2 (MacTel; (J. Gass, 1977)), is a rare degenerative eye disease affecting the macula (Aung, Wickremasinghe, Makeyeva, Robman, & Guymer, 2010; Klein et al., 2010). MacTel is bilateral and progressively affects visual acuity (Charbel Issa et al., 2013/5)(J. D. Gass & Blodi, 1993), reading ability (Finger et al., 2009), and vision-related quality of life (Clemons et al., 2008; Lamoureux et al., 2011). A successful phase II clinical trial has recently been reported for a retinal implant, showing efficacy in slowing disease progression (Chew et al., 2018). However, less invasive, non-surgical and less costly therapies are currently missing. Given the rarity of MacTel and its subtle clinical signs, often requiring several ophthalmological diagnostic tools, the disease has been largely under/misdiagnosed (Charbel Issa et al., 2013/5). Hence insights into the genetic basis of MacTel is key to identify future therapies and to develop predictive models for the disease, allowing better prognostication than currently possible.

We previously published the first genome-wide association study on 476 MacTel patients and 1733 controls (Scerri et al., 2017), identifying and replicating 5 loci. A single nucleotide polymorphism (SNP) at locus 5q14.3 (rs73171800) showed the strongest association with the disease (OR = 2.41) and was previously identified to be associated with the quantitative traits of retinal venular and arterial calibre (Ikram et al., 2010; Sim et al., 2013). The other four loci, 1p12 (OR=1.70, rs477992), 2q34 (OR= 0.50, rs715), 7p11.2 (OR = 1.46, rs4948102), and 3q21.3 (OR = 0.61, rs9820286) were implicated in glycine and serine metabolism (Shin et al., 2014; Xie et al., 2013). Importantly, loci 3q21.3 and 7p11.2 did not reach genome-wide significance and the SNPs identified by Xie et al. at locus 3q21.3 were only in proximity, but not in linkage disequilibrium (LD) with those associated with MacTel. In the same study, with a metabolomics approach we further provided evidence that serum serine, glycine and threonine were depleted in MacTel patients compared to controls. These data provided the first insight into the genetic complexity underpinning MacTel and highlighted the potential involvement of metabolic and vascular trait disturbances in disease aetiology.

Glycine and serine are essential amino acids involved in many fundamental biochemical reactions. Some of the most well known biochemical pathways relying on glycine and serine are the sphingolipids metabolism pathways (KEGG:map00600) and the glycerophospholipids metabolism pathway (KEGG:map00564). Pathogenic variants in genes involved in serine and glycine synthesis lead to severe childhood disorders like phosphoglycerate dehydrogenase deficiency (*PHGDH*, 601815) and glycine encephalopathy (*GLDC*, 605899). Both glycine and serine can be synthesized from one another as well as from other metabolic compounds such as sarcosine and hydroxypyruvate. Additionally, they can both be obtained through dietary intake.

One pathway of interest is the sphingolipids pathway, where serine is required to synthesize sphingomyelin, an abundant lipid in photoreceptor outer segments. Serine deficiency is associated with accumulation of Deoxy Sphingolipids due to alanine being incorporated into palmitoyl-CoA instead of serine. Deoxy Sphingolipids are known to be toxic to different cell

types and in particular neurons (Alecu et al., 2017; Güntert et al., 2016; Penno et al., 2010; Wilson et al., 2018; Zitomer et al., 2009; Zuellig et al., 2014). MacTel patients exhibit elevated serum deoxy-sphingolipids concentrations (Gantner et al 2019), attributable to serine depletion as well as alanine overabundance in MacTel patient serum. Another metabolic pathway of interest is suggested by the previous observation of co-morbidity between Type II diabetes (T2D) and MacTel (Clemons et al., 2013) .

It is still unknown whether traits such as metabolomics abundances, T2D, or retinal vascular calibres have a genuinely causative role on MacTel aetiology. Causation hypothesis is now routinely investigated using Mendelian Randomization (MR). MR exploits the usage of Instrumental Variables (IVs) and, in simple terms, assumes that if a disease is caused by a particular intermediate phenotype, and the latter is caused by genetic variants, then the same genetic variants should be also associated with the disease (Davey Smith & Hemani, 2014; Ebrahim & Davey Smith, 2008; Evans & Davey Smith, 2015). Genetic variants involved in retinal vascular calibre traits and T2D have been previously published (Ikram et al., 2010; Morris et al., 2012; Sim et al., 2013) while the largest metabolomics GWAS meta-analysis to date has recently been published (Lotta et al, in preparation) identifying almost 500 loci that predict the abundance of 142 serum metabolites in humans. The results from these studies can be used to investigate causal drivers of traits such as MacTel susceptibility with the MR approach.

GWA studies on rare complex traits such as MacTel are inevitably underpowered due to the difficulty in recruiting a large number of samples (>500) necessary to interrogate ORs < 2. However, many well-powered GWAS have been performed and are now collated in resources such as the European Bioinformatics Institute GWAS catalogue and methods leveraging such resources to aid underpowered GWA studies have now been developed. An example is Multi-Trait Association GWAS (MTAG) (Turley et al., 2018). MTAG exploits the genetic correlation between traits to prioritise of new loci by leveraging higher powered GWAS from correlated traits to the trait of interest.

Additionally, MacTel patients show substantial heterogeneity of retinal phenotypes, including between pairs of eyes. This, and the lack of clear indicators of early MacTel stages, have frustrated clinical diagnosis. This heterogeneity is likely to be due, at least in part, to different genetic substrates. Apportioning genetic variation to separate retinal malformations has the potential to shed light on the different biological mechanisms via which each locus contributes to the disease. Performing joint genetic and retinal phenotypic data can thus investigate whether different retinal abnormalities are consequences of common or largely independent biological perturbations.

In this study, we test for causality of several traits of interest on MacTel (Figure 1 A). We exploit independent studies on MacTel-related traits to increase discovery power in our disease cohort and both identify new disease loci and resolve already identified ones (Figure 1 B). Then, we use the retinal phenotypic data from MacTel patients to identify key genetic drivers of specific retinal phenotypes (Figure 1 C). Lastly, we collect genetic loci into functional groups by testing their combined effects on ocular phenotypes (Figure 1 D). This study represents the first integrated analysis of MacTel genotypic data with MacTel-related traits as well as retinal imaging phenotypic data. Our study serves as a model for post-GWAS studies with phenotyping datasets.

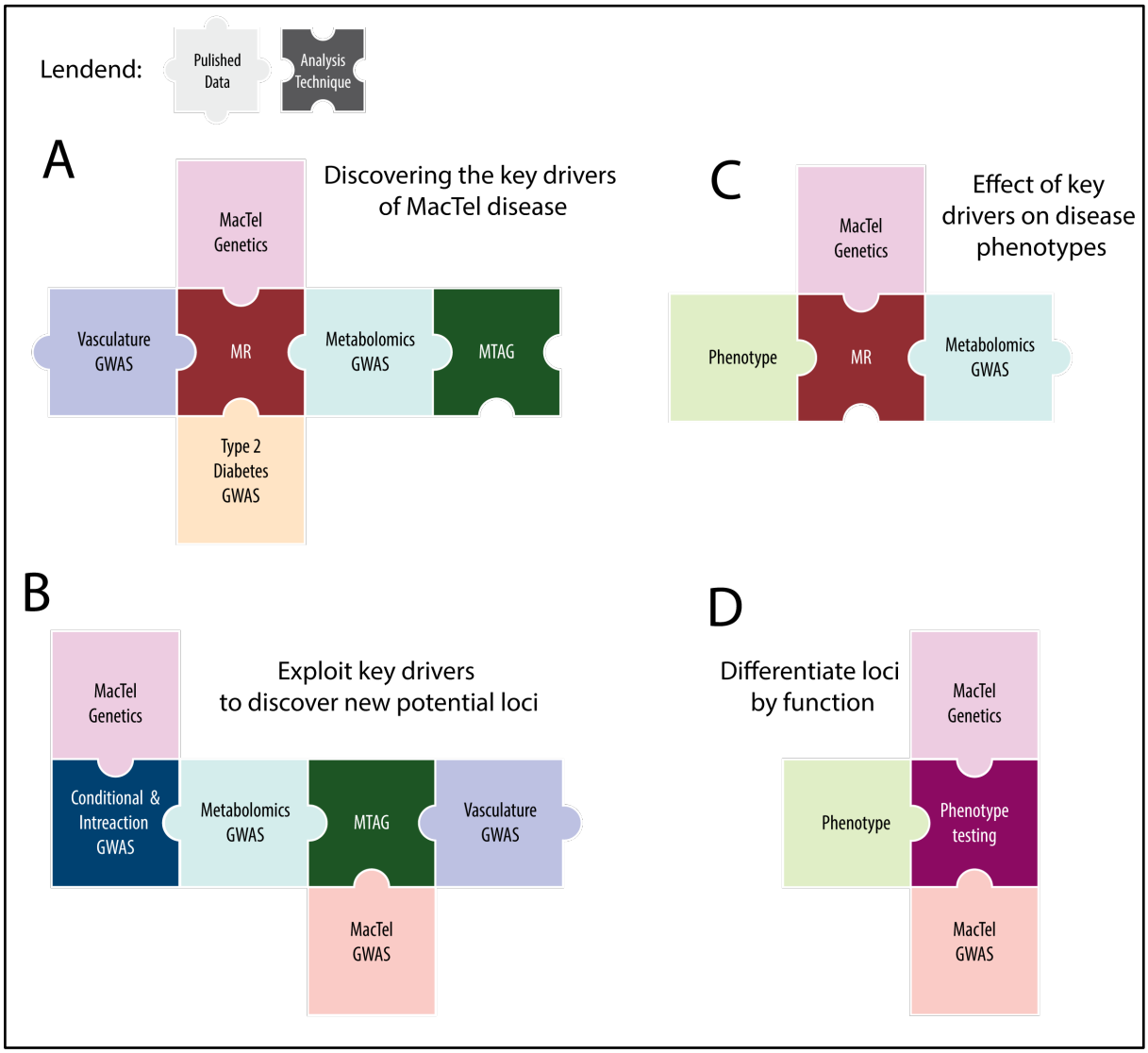


Figure1: Study conceptual map. Each panel represents a study aim and depicts the data and the analysis technique used. Data is presented as pastel color pieces with black writing while analysis techniques are presented as darker panels with white writing. MacTel genetics refers to individual level SNP data available from our previous GWAS study. MacTel GWAS refers to the MacTel GWAS summary statistics. Vasculature GWAS refers to the summary statistics data available from two previous retinal vasculature calibre genetic studies (Ikram et al., 2010; Sim et al., 2013). Metabolomics GWAS refers summary statistics data available from a recently published metabolomics GWAS (Lotta et al). T2D GWAS refers to summary statistics data available from GWAS study on T2D (Morris et al., 2012). MTAG stands for Multi-Trait Association GWAS while MR stands for Mendelian Randomization.

Methods

Unless otherwise stated all statistical and computational analyses were performed using the R statistical software, version 3.5.1. Corrected p-values less than 0.05 were considered statistically significant. For genome-wide association analyses uncorrected p-values< 5e-08 were considered genome-wide significant.

DNA samples and SNP genotyping

Genotypic data was available for 476 MacTel patients and 1733 controls. SNPs were genotyped using the Illumina 5.0M chip as described previously (Scerri et al., 2017). Retinal phenotypic data was available for 455 patients MacTel patients. Genetic predictors (betas or ORs, with corresponding SNP ID and relevant allele) of metabolites and retinal vasculature calibre were provided by the corresponding authors of each study (Lotta et al, (Ikram et al., 2010; Sim et al., 2013)).

Mendelian Randomization Procedures

To perform mendelian randomization analysis with MacTel and metabolites, T2D and retinal vasculature we used the allele-score method for individual-level data as described by Burgess et al. 2016 (Burgess, Dudbridge, & Thompson, 2016). Specifically, we used the weighted score method and estimated Genetically Predicted Metabolites (GPMs) and Genetically Predicted Traits (GPTs), for T2D and retinal vasculature (Table S1). Using a p-value threshold of 5e-08 we extracted SNPs that were significantly associated with metabolites or traits (retinal venular calibre, retinal arteriolar calibre, and T2D). We constructed GPMs and GPTs using the magnitudes of the significant SNPs as weights in the manner of a Polygenic Risk Score (Supplementary Methods). We used logistic regression models to test for association between each GPM and GPT and MacTel susceptibility, correcting for genetically determined sex at birth and the first principal component, as in our previous publication (Scerri et al., 2017) to account for batch effects including population stratification. We used Benjamini-Hochberg multiple testing correction to account for the false discovery rate. A sequential conditional modelling approach was used to identify GPMs independently associated with the disease. Specifically, the most significant GPMs were iteratively added as covariates in the logistic regression model and re-tested, until no GPM was nominally significant after FDR correction.

Conditional and Interaction GWAS and MTAG analysis

The conditional GWA studies were performed by including GPMs or GPTs in the logistic models as covariates. For the interaction GWA studies, an interaction term was included between each SNP and the GPM or GPT of interest. Consistent with the original GWAS analysis on this data, each model also included genetically determined sex at birth and the first principal component (Scerri et al., 2017). Analyses were performed using Plink v1.9 (Purcell et al., 2007). Results from the conditional and interaction GWAS were analysed using FUMA (Watanabe, Taskesen, van Bochoven, & Posthuma, 2017). LocusZoom plots were produced by using the LocusZoom software (Pruim et al., 2010). MTAG analysis (Turley et al., 2018) was performed by integrating the available summary statistics from the same study as those used to perform mendelian randomization for each trait of interest using the MTAG software.

Retinal phenotype clustering

As part of the Natural History Study and Registry of Macular Telangiectasia (Clemons et al., 2010) longitudinal retinal phenotypic data (colour fundus images, optical coherence tomography and fluorescein angiography images) was collected over the years 2005-2015 from 1,716 patients (3,410 eyes in total). The data consisted of 143 spatial measurements of retinal phenotypes. A list of all phenotypes and measurement methods is provided in Table S2. A detailed description of the corresponding methods can be found elsewhere (Clemons et al., 2010). Each phenotype was measured in 9 different subfields of the retina (defined by the ETDRS grid (Mathew et al., 2013) Figure S1). The phenotype data was cleaned by performing

missing data imputation (Supplementary Methods). The cleaned dataset contained 119 phenotypes that were collapsed into 30 biologically relevant endophenotypes using factorial analysis as described in detail in Supplementary Methods. With this technique, we were able to distil the set of observed macular phenotypes into a much smaller set of endophenotypes encapsulating the key features of clinical diagnostics of MacTel Figure S6. An example endophenotype arose from leakage measured in multiple discrete retinal areas leading to a common 'leakage' endophenotype (Figure S6-B).

Investigating the relationship of retinal endophenotypes with genetic drivers

We tested for association between retinal endophenotypes with significant disease-associated GPMs, GPTs and prioritised SNPs. The dataset containing retinal endophenotypes and genetic information included 3,280 observations from 455 MacTel patients (907 eyes in total) with an average of 3.6 observations per eye over 10 years. Association testing was performed using a linear mixed model approach, assuming an additive effect of SNP alleles that increased MacTel risk. P-values were corrected using an adaptive Benjamini-Hochberg procedure (Benjamini, Krieger, & Yekutieli, 2006). Further details are provided in Supplementary Methods.

Results

Discovering key drivers of MacTel disease: Metabolites

The MR procedure (Figure 1A) applied to the metabolite panel highlighted 14 GP metabolites significantly associated with MacTel Table 1. A list of all metabolite associations is provided in Table S3. The two most highly significant metabolites were GP serine (OR = 0.52, p = 8.38e-28) and GP glycine (OR = 0.56, p = 5.44e-19). Interestingly, in contrast to the metabolomics results from our previous study, GP threonine did not reach significance (OR = 0.98, p = 0.996). Other significant metabolites were Phosphatidylcholine diacyl C32:1 (OR=1.23, p=0.0025), lysoPhosphatidylcholine acyl C14:0, Phosphatidylcholine diacyl C34:1 (OR=1.23, p=0.0025), Phosphatidylcholine with acyl-alkyl residue C38:1 (OR=1.23, p=0.0025), Arginine (OR=1.18, p=0.0283) and Phenylalanine (OR=0.84, p=0.0285). GP alanine was borderline significant (OR = 1.55, p = 0.069)

The significant GPMs correlated with each other in our sample (Figure S2), mirroring correlations observed with directly measured metabolomics abundances. When all GPMs were re-tested in a model including GP serine as a covariate, GP glycine was the only associated metabolite (OR = 0.76, p = 2.2e-02) suggesting that the effect of GP glycine on MacTel is largely due to biochemical interactions with serine. GP serine remained significant when included alongside glycine (OR = 0.6, p = 3.3e-11). No other GP metabolites remained significant once GP serine and glycine had been included (Table S3).

Metabolite	Beta	OR	P-value	Family
Serine	-0.651	0.521	8.38E-28	Amino Acids
Glycine	-0.576	0.562	5.44E-19	Amino Acids
Phosphatidylcholine diacyl C32:1	0.208	1.231	0.0025	Glycerophospholipids
lysoPhosphatidylcholine acyl C14:0	0.208	1.231	0.0025	Glycerophospholipids
Phosphatidylcholine diacyl C34:1	0.208	1.231	0.0025	Glycerophospholipids
Phosphatidylcholine with acyl-alkyl residue C38:1	0.202	1.224	0.0030	Glycerophospholipids
Arginine	0.168	1.183	0.0283	Amino Acids
Phenylalanine	-0.166	0.847	0.0285	Amino Acids

Table 1: Significant associations between genetically predicted metabolites (GPMs) and MacTel. Regression coefficients are presented in the “Beta” column. Odds Ratios (OR) are relative to a single standard deviation increase on the GPM scale.

The GWAS used to predict the genetic complement of serine was largely underpowered compared to the glycine GWAS. Hence, we attempted to boost serine GWAS results by combining them with the glycine GWAS results in an MTAG analysis. By using the newly identified serine loci derived by its biochemically-based genetic correlation with glycine we constructed a new and more powerful genetic predictor of serine (GP MTAG serine, Table S1). We found GP MTAG serine sufficiently encompassing all the metabolic predictive signal for MacTel as the original GP glycine was no longer required in the model once GP MTAG serine was included (Figure 1A). Further MR testing of an even broader set of 248 GP metabolite abundances using results from Shin et al study (“metabolomics gwas server,” n.d.; Shin et al., 2014) (Table S1) found few significant contributions from other metabolites although none remained significant after inclusion of GP MTAG serine in the model (results of MTAG and MR from Shin et al metabolites available in Supplementary Materials and Table S4).

Discovering key drivers of MacTel disease: T2D and retinal vasculature

Genetically predicted T2D was significantly associated with the disease (OR = 1.96, p = 0.0008). When tested separately, both GP arteriolar (OR = 1.38, p = 2.22e-10) and GP venular

calibre (OR = 1.19, p = 0.0007) were also significantly associated. However, we found that this result was driven by the association with two single SNPs: rs2194025 (used in predicting arteriolar calibre) and SNP rs17421627 (used in predicting venular calibre), both in strong LD with SNP rs73171800 at locus 5q14.3, identified in our original GWAS (OR = 2.41, p = 7.7e-17 (Scerri et al., 2017)). When including both vascular traits in the same model, only arteriolar calibre remained significant, due to the reduced number of SNPs (N = 2) used to construct that GPT (Figure S3). When GPTs for retinal arteriolar calibre and type 2 diabetes were included in a model with GP glycine and serine, all remained significant.

Additionally, we visually compared the effect sizes for each SNP used to define GPMs and GPTs on both MacTel and the metabolites or trait. Such comparison is displayed in Figure 2. As expected, a negative relationship could be detected between the effect on metabolic abundance and MacTel risk for the SNPs used to define glycine and serine. A positive effect was observed for T2D and vasculature traits, while no relationship was detected for threonine. We additionally observed that the regression line for all significant traits except from the vasculature related traits did not have intercept terms that were significantly different from 0 ($b_{0,serine}$= –0.07, p=0.2 ; $b_{0,glycine}$=0.070, p=0.131 ; $b_{0,T2D}$=0.002, p=0.88) suggesting a non-pleiotropic effect of the instruments used to test such traits (Bowden, Davey Smith, & Burgess, 2015).

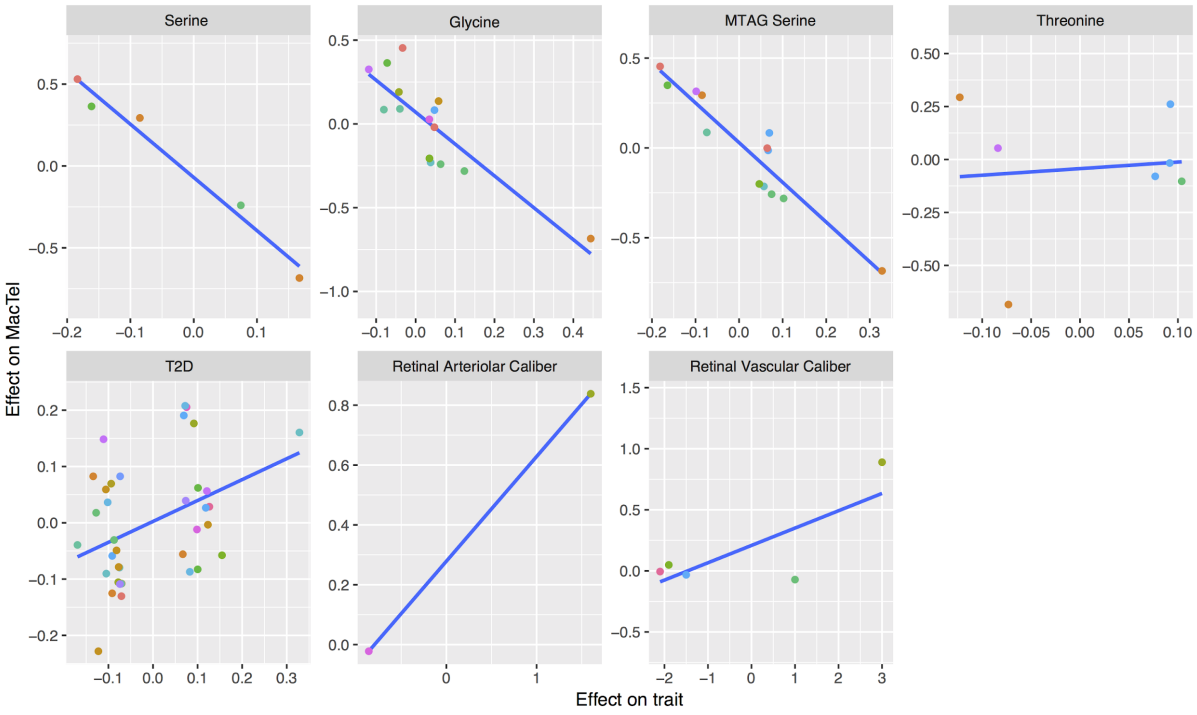


Figure 2: Effect size comparisons between MacTel and specific trait of SNPs used to define GPM and GPT. In this plot, each dot represents a SNP. Traits are divided into panels and each panel contains the SNPs that were found to have a genome-wide significant effect on the specific metabolite or trait. The x-axis captures the effect sizes of the SNPs on the trait while the y-axis captures the effect of the same SNPs on MacTel risk. The blue line represents a simple regression line between the two. Metabolites or traits causally affecting the disease are expected to have correlated effect sizes. Colour of the dots represents chromosomal positions where different chromosomes are represented by different colours.

Using causal traits to discover new disease susceptibility loci for MacTel

MTAG results

To discover new loci putatively involved in MacTel, we performed MTAG analysis by including GWAS summary results for glycine, serine, retinal venular calibre, and retinal arteriolar calibre along-side MacTel disease status (Figure 1B). T2D was not included, as full genome-wide summary statistics from the Morris et al 2012 Nat Genetics study (Morris et al., 2012)were not available to us. The number of SNPs shared across all datasets was 799,436. This analysis confirmed the three previously identified loci (Scerri et al., 2017) and revealed 5 novel loci attaining genome-wide significance Table 2. A detailed table of results output by the FUMA software on these results is available In Table S5.

Rsid	Locus	MTAG P-val MacTel	Novel GW Sig.	P-val Art	P-val Ven	P-val Gly	P-val Ser	P-val MacTel
				Sim et al 2013	Ikram et al 2010	Lotta et al 2019	Lotta et al 2019	Scerri et al 2017
rs3815630	2q34	**7.62E-143**	No	*1.51E-01*	*6.23E-01*	**2.23E-308**	**6.50E-31**	**4.43E-08**
rs1035387	5q14.3	**1.16E-16**	No	**1.50E-09**	**1.15E-08**	*5.43E-01*	*7.53E-01*	**1.16E-16**
rs477992	1p12	**2.21E-15**	No	*7.44E-01*	*7.95E-04*	**6.09E-07**	**2.01E-80**	**2.60E-12**
rs4543497	7p11.2	**2.22E-15**	Yes	*3.56E-01*	*2.07E-01*	**9.01E-28**	**3.29E-54**	8.11E-06
rs4841132	8p23.1	**8.43E-14**	Yes	*5.55E-02*	*2.95E-02*	**1.73E-34**	1.23E-04	4.54E-02
rs1563075	16q23	**1.79E-13**	Yes	*6.86E-01*	*3.77E-01*	**2.57E-28**	1.19E-01	4.66E-03
rs2954021	8q24.1	**1.53E-12**	Yes	*3.89E-01*	*6.33E-01*	**2.35E-21**	**2.40E-10**	8.70E-05
rs4742212	9p24	**1.19E-09**	Yes	*4.73E-01*	*6.88E-01*	**3.06E-34**	1.69E-04	2.82E-01

Table 2: Genome-wide significant loci for MacTel from MTAG analysis. MacTel GWAS results were combined with GWAS results from retinal venular (Ven) and arteriolar (Art) calibre, and serine (Ser) and glycine (Gly) abundance. Bold text indicates genome-wide significant P-values on respective original studies. Grey italicised p-values are associations expected to be null a priori.

The three confirmed loci were locus 2q34 (rs3815630), 1p12 (rs477992) and 5q14.3 (rs1035387). Among the new genome-wide significant loci is locus 7p11.2 (rs4543497, p = 2.22e-15). In our original MacTel GWAS this locus did not reach genome wide significance (p = 8.11e-06) but was deemed important given its established relationship with glycine and serine. The other four novel loci from the MTAG analysis were: 8p23.1 (rs4841132, p = 8.43e-

14), 16q23 (rs1563075, p = 1.79e-13), 8q24.1 (rs2954021, p = 1.53e-12), and 9p24 (rs4742212, p = 1.19e-09). We did not observe any significant MTAG result for locus 3q21.3 (rs9820286), originally proposed to affect MacTel though a connection with glycine and serine (Scerri et al., 2017). This SNP was in fact not genome-wide significant for either glycine (p = 0.04063) or serine (p = 0.6526).

Conditional and Interaction GWAS

To uncover genetic correlates of MacTel independent of glycine and serine, we performed a GWAS conditioning on these genetically predicted traits (Figure 1B). This analysis identified four genome-wide significant peaks (Figure S4 a).

As expected, the original signal on locus 5q14.3 (rs73171800) remained significant, confirming the independence of this locus from genetic drivers of serine and glycine. A second genome-wide significant signal on locus 3p24.1 (rs35356316, p = 3.10e-08) was also identified, which did not reach significance in our original study (p=1.54e-07). It is situated in a 'gene desert' proximal to the genes *EOMES* and *SLC4A4*. Another SNP in very close proximity reached genome-wide significance as in the original MacTel GWAS study (Scerri et al., 2017) but was believed to be a false positive, given the lack of LD with any other significant SNP (Figure S4 b).

The remaining two conditionally significant SNPs tagged independent signals in locus 19p13.2 (Figure S4 c-d). SNP rs36259 is an exonic non-synonymous SNP located in the *CERS4* gene which achieved close to genome-wide significance (p=6.270e-08) and was nominally significant in our original study (p = 1.69e-7). In close proximity, we found an independent intergenic SNP rs4804075 (p = 3.72e-07) for which we found no evidence for an eQTL effect on any neighbouring genes and which did not reach genome-wide significance. These results remain significant when conditioning on SNP rs73171800 (locus 5q14.3) and GP T2D (results not shown).

We performed additional GWAS analyses testing for interactions between all SNPs with GP serine, GP glycine, GP T2D, and SNP rs73171800 (Figure 1B). However, no further significant interacting loci were found(Figure S5 a-d).

Effect of key drivers on retinal endophenotypes

Given the small sample size of the endophenotypic data (N = 455), we decided to only test for association between endophenotypes and GP glycine and serine, as these bore the clearest association with disease aetiology (Figure 1C). No significant association between GP glycine and endophenotypes were found when accounting for GP serine, and we therefore excluded it from additional testing. High GP serine levels were found to be protective against loss of retinal transparency (b = -0.54, p = 0.013), leakage at the retinal pigment epithelium (RPE) in progression areas (b = -0.43, p = 0.036) and suggestively in MacTel area (b = -0.28, p = 0.055), and leakage at the outer capillary network in progression areas (b = -0.40, p = 0.046). We also found suggestive significance for protection from irregularities of the RPE leakage in

the MacTel area (b = -0.38, p = 0.060) and progression area (b = -0.26, p = 0.089). Higher GP serine levels were associated with protection from macular thinning in the MacTel area (b = -0.44, p = 0.036), inferior inner area (b = -0.4, p = 0.036), and suggestive protection on nasal inner area (b = -0.32, p = 0.054), and foveal area (b = -0.42, p = 0.051).

Together these results highlighted that not only GP serine was associated with the disease risk but was, more specifically, also associated with endophenotypes characterising disease progression.

Determining the effects of GWAS loci on endophenotypes

To determine the likely functional impacts of the original GWAS loci we tested them for association with retinal endophenotypes (Figure 1D and Figure 3).

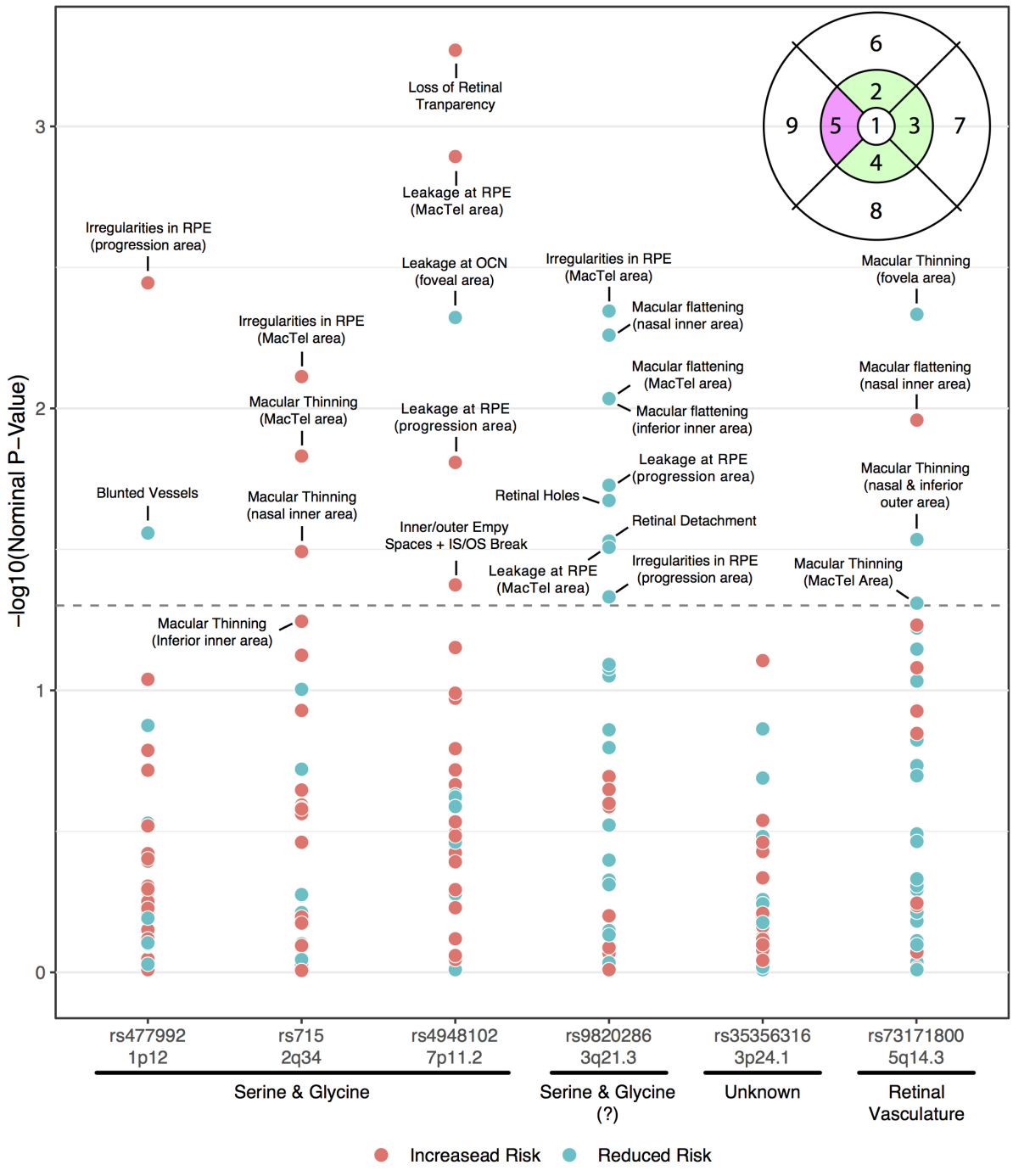


Figure 3: Association plot summarising all nominally significant association between MacTel SNPs and endophenotypes. SNPs are divided into categories based on MR results. Each dot represents the corrected p-value for each association between all SNPs and endophenotypes. Red dots indicate that MacTel risk alleles for that SNP also increase the risk of the retinal phenotype, while blue dots indicate a reduced risk. The top right corner displays the ETDRS grid for the right eye (OD) (Mathew et al., 2013) used to divide the endophenotypes into 9 retinal areas; 1: foveal area, 2: superior inner area, 3: nasal inner area, 4: inferior inner area, 5: temporal inner area, 6: superior outer area, 7: nasal outer area, 8: inferior outer area, and 9: temporal outer area. The pink area is the MacTel area while the green areas are progression areas.

Unsurprisingly, in this analysis no association remained significant after correction for multiple testing, highlighting the extremely modest effect sizes of single SNPs on disease endophenotypes.

However, using the nominally uncorrected p-values ($p^* < 0.05$) in an explorative approach we observed that locus 1p12 (rs477992) was associated with an increased risk of irregularities of the RPE in the progression area ($b = 0.12$, $p^* = 0.003$), but protected against risk of blunted vessels ($b = -0.08$, $p^* = 0.028$).

Locus 2q34 (rs715) increased the risk of macular thinning in the MacTel area ($b = 0.16$, $p^* = 0.014$) and nasal inner area ($b = 0.12$, $p^* = 0.032$) and the risk of irregularities in the RPE layer in the MacTel area ($b = 0.19$, $p^* = 0.007$).

The MacTel risk allele at locus 7p11.2 (rs4948102) increased the risk of retinal transparency ($b = 0.17$, $p = 0.0005$), leakage at the RPE in both the MacTel area ($b = 0.13$, $p^* = 0.001$) and progression area ($b = 0.12$, $p^* = 0.015$), as well as presence of inner empty spaces and IS/OS break ($b = 0.10$, $p^* = 0.043$). This locus however protected against leakage in the outer layer of the foveal area ($b = -0.13$, $p^* = 0.004$).

Together, the three loci thought to be underpinned by genetic variants acting on genes important in the glycine/serine metabolism (1p12/rs477992/*PHGDH*, 2q34/rs715/*CPS1*, and 7p11.2/rs4948102/*PSPH*) shared an overall susceptibility increment to RPE endophenotypes and other macular phenotype risks.

Interestingly the MacTel risk allele at locus 3q21.3 (rs9820286) was associated only with protection from retinal disease endophenotypes. Specifically, locus 3q21.3 was associated with protection against retinal holes ($b = -0.08$, $p^* = 0.021$), retinal detachment ($b = -0.12$, $p = 0.030$), irregularities at the RPE layer in the MacTel area ($b = -0.23$, $p^* = 0.005$) and progression areas ($b = -0.12$, $p^* = 0.047$); leakage at the RPE level in the MacTel area ($b = -0.12$, $p^* = 0.031$) and progression area ($b = -0.17$, $p^* = 0.019$); and foveal slope flattening in MacTel area ($b = -0.17$, $p^* = 0.009$), inferior inner area ($b = -0.17$, $p^* = 0.009$), and nasal inner area ($b = -0.19$, $p^* = 0.006$).

Locus 5q14.3 (rs73171800) corresponded to protection against macular thinning in the foveal area ($b = -0.19$, $p^* = 0.005$) and nasal-inferior outer areas ($b = -0.14$, $p^* = 0.030$) and increased risk of foveal flattening in the nasal inner area ($b = 0.15$, $p^* = 0.011$). The effect of SNP rs73171800 on macular thickness was consistent with recent findings (Gao, Huang, & Kim, 2018). We did not find any nominally significant association between SNP rs35356316 located in locus 3p24.1 and the aggregated retinal endophenotypes.

Additionally, we attempted to cluster the loci with hierarchical clustering performed on their regression coefficients (Figure S7). This analysis highlighted the unique endophenotype effect profiles of the 5q14.3 (rs73171800) and 3q21.3 (rs9820286) loci, and their distinctiveness in comparison to the other three loci, known to act on the serine/glycine pathway.

Discussion

This study exploited well-powered publicly available GWAS data from traits of interest as well as deep phenotyping data to further investigate the genetic aaetiology of MacTel, a moderately rare retinal disorder on which the first GWAS was only performed in 2017. These analyses have discovered causal genetic drivers for MacTel, delineating metabolic independent genetic contributors to the disease, and quantifying the contribution of each trait and locus to retinal degeneration.

Mendelian randomization and publicly available data were used to “genetically predict” abundances of different metabolites. Even though all GPMs in this study were generated using estimated SNP allele dosage effects from blood-based metabolomic studies rather than retina, we believe that our results highlight that this is a highly relevant approach useful for other retinal disorders.

We found genetically-predicted serine depletion as the strongest causal driver of MacTel risk, with an association magnitude much greater than any other single SNP. This pronounced effect reflects the power of combining multiple genetic variants associated with the disease through a shared biological mechanism; and further emphasizes the causative role of serine in this disease. The direction of association between GP serine and MacTel agrees with the one of direct serum serine measured in our previously published study. This act as compelling additional evidence for this causal association. Indeed, the disease odds of MacTel doubled for every standard unit decrease in GP serine and individuals in the lowest 20% of GP serine presented an OR of 6.34 for MacTel compared to individuals in the top 20%. Our study results also confirm the pronounced role that serine plays in disease progression apart from its aetiology. As this analysis was performed only among MacTel patients the heterogeneity of retinal phenotypes at early disease stages may be under-represented. GP serine was nevertheless able to partly discriminate between subjects with more advanced retinal abnormalities from those without. For example, lower GP serine had a clear association with retinal greying in all retinal areas, which is not observed in all MacTel patients (Charbel Issa et al., 2013/5). GP serine also affected retinal thinning in the temporal parafoveal area, a marker of photoreceptor degeneration. However, although our results suggest a highly plausible biochemical explanation for differences in disease heterogeneity and progression, we acknowledge that our study used only aggregated retinal phenotypes and proxy measures of progression. Longitudinal data documenting progression phenotype is now required to confirm these findings. High concentrations of deoxy-sphingolipids, a byproduct of serine deficiency, have recently been shown to cause MacTel (Gantner et al 2019). Our results provide evidence of causal genetically-encoded serine depletion, which likely contributes to the disease by promoting deoxy-sphingolipid biosynthesis.

Our results indicate that the association between glycine and MacTel is likely an artefact of the shared genetic signal between this metabolite and serine. Glycine may have emerged as a highly associated metabolite instead of serine due to the considerably greater sample size used to construct the genetic predictor of the former. We failed to find any effect of glycine independent of serine in our phenotypic analyses. Given these results, it is possible that genetic depletion of glycine results in lower serine abundance as the latter can be biosynthesized by the former. This, in turn, could be causative for MacTel via heightened deoxy-sphingolipid synthesis.

Threonine was the second most differentially abundant metabolite in MacTel patient serum(Scerri et al., 2017). However, GP threonine was not statistically associated with the disease, indicating that lower levels of threonine do not have a direct causative role on MacTel. Rather, as for glycine, this result may be due to biochemical interactions between threonine and serine as indicated by the multiple biochemical connections of these metabolites in glycine, serine and threonine metabolic pathway (KEGG:map00260).

By using Multi-Trait Association GWAS (MTAG) we confirmed the central role of locus 7p11.2 (rs4543497), which was previously found to be only nominally significant, and identified a new locus, 8q24.1 (rs2954021). The latter locus is of great interest since rs2954021 is predictive of endogenous serum serine concentrations (Lotta et al) and thus might confer MacTel risk by contributing to the same serine pathway as previously identified SNPs. We did not find any new signal arising from shared genetic correlation with retinal vascular calibre traits.

Although not significant after correction for multiple testing, we observed a pronounced effect of SNP rs73171800 on macular thickness, specifically, a protective effect of C allele against macular thinning, which is tantamount to an increase in macular thickness. Interestingly, a large GWAS study, performed on 68,423 subjects in the UK Biobank database found that the G allele of SNP rs17421627 - corresponding to the C allele of SNP rs73171800 (LD with rs73171800 r^2=0.67), was significantly associated with increased macular thickness(Gao et al., 2018). Further, a targeted study of SNP rs17421627, leveraged the deep conservation of this locus to demonstrate enhancer activity which modifies the retinal vasculature. Madelaine et al substituted the homologous zebrafish locus with a construct containing rs17421627, and showed enhancer activity and changed expression of a proximal microRNA, mir-9-2 (homologous with human mir-9-5) (Madelaine et al., 2018). Both mir-9-2 knock-down and endogenous enhancer knock-out animals showed dysmorphic retinal vasculature, indicating that rs17421627 may act on miRNA expression to modify the formation of the retinal vasculature in humans(Gao et al., 2018) Although our MR analysis did not find a causal relationship between retinal vasculature calibre and MacTel, it may be that other features of the vasculature, for example, leakage, branching or integrity, are modified in rs17421627 carriers, and account for the gross differences in macular thickness. Further deep phenotyping of the retinal vasculature in these individuals may yield the physiological basis for this phenotype and its impact on MacTel.

By conditioning on GP serine, glycine and T2D we revealed genetic signals that are contributing to MacTel independent of these. Among these was a new genome-wide significant locus 3p24.1, tagged by SNP rs35356316. This SNP lies between the genes *EOMES* (604615) and *SLC4A4* (603345). Interestingly, *EOMES* encodes a transcriptional activator which is shown in mouse studies to interact with the retinal transcription factor Pou4f2 (Mao et al., 2008). The downstream gene *SLC4A4* encodes a sodium-coupled bicarbonate transporter which is expressed in Müller glia and the RPE, and functions to balance pH in the subretinal space (Pushkin et al., 1999).

Conditional analysis also revealed a glycine/serine/T2D independent MacTel signal at locus 19p13.2. The SNP rs36259 which tags this locus is an exonic, non-synonymous SNP located within the gene *CERS4* (615334). This gene encodes a dihydroceramide synthase, which operates in the sphingolipid biosynthesis pathway, downstream of SPTLC (Riebeling, Allegood, Wang, Merrill, & Futerman, 2003). Just as serine depletion is associated with

defective sphingolipid synthesis, this locus may reduce sphingolipid production and thus contribute to MacTel.

Type 2 diabetics are over-represented among MacTel patients (Clemons et al., 2013) and we found a weak positive association between GP T2D and MacTel risk. A possible explanation for this is that T2D involves major perturbations of patient metabolism. Specifically, a recent meta-analysis of metabolite abundances associated with pre-diabetes and/or T2D, found that glycine depletion tends to occur in this disease (Guasch-Ferré et al., 2016). We show that glycine depletion is not likely to be causative for MacTel and is likely to be a consequence of metabolomics co-regularisation counterbalancing the genetically induced serine depletion. We thus speculate that diabetes may be a consequence rather than a cause of the metabolic phenotype underlying MacTel. Methods such as bi-directional MR address the question of directionality of effect when two traits are closely related. However, caution is advised when using these techniques with relatively weak prior genetic knowledge (Davey Smith & Hemani, 2014). We recognise that although knowledge needed to perform such an experiment may be available for T2D, this is not the case in MacTel for which currently identified genetic signals have only modest explanatory power. Indeed, T2D risk remained significant even after controlling for GP serine and GP glycine, which indicates the possibility of a separate mechanism unrelated to glycine/serine metabolism.

Our analysis revealed that locus 3q21.3, which was previously believed to act on the disease by modifying serum glycine and serine, is instead likely disconnected from these metabolites, as is the case for loci 3p24.1 and 5q14.3. Furthermore, these three loci do not cluster with each other, implying that additional potential disease mechanisms exist, beyond the already strongly implicated glycine/serine metabolic dysregulation.

To conclude, by integrating different sources of genetic and phenotypic data related to macular telangiectasia type II with advanced statistical methods, we find evidence that genetically-encoded serine depletion causes disease onset and likely also progression. The role of glycine, threonine and type 2 diabetes, as well as other vascular traits, is diminished in our results. We apportion genetic signals to different retinal phenotypes, and present evidence of independent contributions from genetic variants potentially related to vasculature, pH regulation and sphingolipid synthesis. We look forward to further targeted studies of disease progression and experimental validation of serine-independent contributions, to further resolve the complex aaetiology of this disease and enhance treatment efficacy. Our in-depth analysis of deep phenotyping available from our patient cohort, in addition to the usage of several recently made available GWAS datasets serve as a model as to how such resources can be used to dissect GWAS results, especially for rarer traits.

Acknowledgements:

We would like to acknowledge the funding support from the Lowy Medical Research Institute. This work was also made possible through the Victorian State Government Operational Infrastructure Support and Australian Government National Health and Medical Research Council (NHMRC) independent research Institute Infrastructure Support Scheme (IRIISS). RB was supported by the Melbourne International Research Scholarship. BREA was supported by an NHMRC early career Fellowship (1157776). MB was supported by an NHMRC Senior Research Fellowship (1102971) and Program Grant (1054618). We are

grateful to Dr Saskia Freytag, Dr Anna Quaglieri, Prof Terry Speed, Prof Gordon Smyth, Dr Mari Gantner, Dr Kevin Eade, Dr Martina Wallace, A/Prof Christian Metallo, Prof Martin Friedlander, Dr Tjebo Heeren, Dr Mali Okada, Dr Sasha Woods, Prof Marcus Fruttiger, and Dr Catherine Egan for their extremely helpful contribution that greatly improved the quality of this manuscript. We are especially grateful to Dr Xueling Sim and Prof Wong Tien Yin for kindly sharing their summary data on retinal vascular calibre traits.

Declaration of Interests:

The authors declare no conflict of interest.

Web Resources:

FUMA: http://fuma.ctglab.nl/
Locus zoom: http://locuszoom.org/
Metabolomics GWAS server: http://metabolomics.helmholtz-muenchen.de/gwas/

References:

Alecu, I., Tedeschi, A., Behler, N., Wunderling, K., Lamberz, C., Lauterbach, M. A. R., … Penno, A. (2017). Localization of 1-deoxysphingolipids to mitochondria induces mitochondrial dysfunction. *Journal of Lipid Research*, *58*(1), 42–59.

Aung, K. Z., Wickremasinghe, S. S., Makeyeva, G., Robman, L., & Guymer, R. H. (2010). The prevalence estimates of macular telangiectasia type 2: the Melbourne Collaborative Cohort Study. *Retina* , *30*(3), 473–478.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507.

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, *44*(2), 512–525.

Burgess, S., Dudbridge, F., & Thompson, S. G. (2016). Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in Medicine*, *35*(11), 1880–1906.

Charbel Issa, P., Gillies, M. C., Chew, E. Y., Bird, A. C., Heeren, T. F. C., Peto, T., … Scholl, H. P. N. (2013/5). Macular telangiectasia type 2. *Progress in Retinal and Eye Research*, *34*, 49–77.

Chew, E. Y., Clemons, T. E., Jaffe, G. J., Johnson, C. A., Farsiu, S., Lad, E. M., … Macular Telangiectasia Type 2-Phase 2 CNTF Research Group. (2018). Effect of Ciliary Neurotrophic Factor on Retinal Neurodegeneration in Patients with Macular Telangiectasia Type 2: A Randomized Clinical Trial. *Ophthalmology*. https://doi.org/10.1016/j.ophtha.2018.09.041

Clemons, T. E., Gillies, M. C., Chew, E. Y., Bird, A. C., Peto, T., Figueroa, M. J., … MacTel Research Group. (2010). Baseline characteristics of participants in the natural history study of macular telangiectasia (MacTel) MacTel Project Report No. 2. *Ophthalmic Epidemiology*, *17*(1), 66–73.

Clemons, T. E., Gillies, M. C., Chew, E. Y., Bird, A. C., Peto, T., Figueroa, M., … Macular Telangiectasia Research Group. (2008). The National Eye Institute Visual Function Questionnaire in the Macular Telangiectasia (MacTel) Project. *Investigative Ophthalmology & Visual Science*, *49*(10), 4340–4346.

Clemons, T. E., Gillies, M. C., Chew, E. Y., Bird, A. C., Peto, T., Wang, J. J., … Macular Telangiectasia Project Research Group. (2013). Medical characteristics of patients with macular telangiectasia type 2 (MacTel Type 2) MacTel project report no. 3. *Ophthalmic Epidemiology*, *20*(2), 109–113.

Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, *23*(R1), R89–R98.

Ebrahim, S., & Davey Smith, G. (2008). Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics*, *123*(1), 15–33.

Evans, D. M., & Davey Smith, G. (2015). Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. *Annual Review of Genomics and Human Genetics*, *16*, 327–350.

Finger, R. P., Charbel Issa, P., Fimmers, R., Holz, F. G., Rubin, G. S., & Scholl, H. P. N. (2009). Reading performance is reduced by parafoveal scotomas in patients with

macular telangiectasia type 2. *Investigative Ophthalmology & Visual Science*, *50*(3), 1366–1370.

Gao, X. R., Huang, H., & Kim, H. (2018). Genome-wide association analyses identify 139 loci associated with macular thickness in the UK Biobank cohort. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddy422

Gass, J. (1977). Some problems in the diagnosis of macular diseases. *Symposium on Retinal Diseases*, 268–270.

Gass, J. D., & Blodi, B. A. (1993). Idiopathic juxtafoveolar retinal telangiectasis. Update of classification and follow-up study. *Ophthalmology*, *100*(10), 1536–1546.

Guasch-Ferré, M., Hruby, A., Toledo, E., Clish, C. B., Martínez-González, M. A., Salas-Salvadó, J., & Hu, F. B. (2016). Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. *Diabetes Care*, *39*(5), 833–846.

Güntert, T., Hänggi, P., Othman, A., Suriyanarayanan, S., Sonda, S., Zuellig, R. A., … Ogunshola, O. O. (2016). 1-Deoxysphingolipid-induced neurotoxicity involves N-methyl-d-aspartate receptor signaling. *Neuropharmacology*, *110*(Pt A), 211–222.

Ikram, M. K., Sim, X., Xueling, S., Jensen, R. A., Cotch, M. F., Hewitt, A. W., … Wong, T. Y. (2010). Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo. *PLoS Genetics*, *6*(10), e1001184.

Klein, R., Blodi, B. A., Meuer, S. M., Myers, C. E., Chew, E. Y., & Klein, B. E. K. (2010). The prevalence of macular telangiectasia type 2 in the Beaver Dam eye study. *American Journal of Ophthalmology*, *150*(1), 55–62.e2.

Lamoureux, E. L., Maxwell, R. M., Marella, M., Dirani, M., Fenwick, E., & Guymer, R. H. (2011). The longitudinal impact of macular telangiectasia (MacTel) type 2 on vision-related quality of life. *Investigative Ophthalmology & Visual Science*, *52*(5), 2520–2524.

Madelaine, R., Notwell, J. H., Skariah, G., Halluin, C., Chen, C. C., Bejerano, G., & Mourrain, P. (2018). A screen for deeply conserved non-coding GWAS SNPs uncovers a MIR-9-2 functional mutation associated to retinal vasculature defects in human. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky166

Mao, C.-A., Kiyama, T., Pan, P., Furuta, Y., Hadjantonakis, A.-K., & Klein, W. H. (2008). Eomesodermin, a target gene of Pou4f2, is required for retinal ganglion cell and optic nerve development in the mouse. *Development* , *135*(2), 271–280.

Mathew, R., Sivaprasad, S., Florea, D., Leung, I., Sallo, F., Clemons, T., … Peto, T. (2013). Agreement between time-domain and spectral-domain optical coherence tomography in the assessment of macular thickness in patients with idiopathic macular telangiectasia type 2. *Ophthalmologica. Journal International D'ophtalmologie. International Journal of Ophthalmology. Zeitschrift Fur Augenheilkunde*, *230*(3), 144–150.

metabolomics gwas server. (n.d.). Retrieved December 4, 2018, from http://metabolomics.helmholtz-muenchen.de/gwas/

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., … DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, *44*(9), 981–990.

Penno, A., Reilly, M. M., Houlden, H., Laurá, M., Rentsch, K., Niederkofler, V., … Hornemann, T. (2010). Hereditary sensory neuropathy type 1 is caused by the accumulation of two neurotoxic sphingolipids. *The Journal of Biological Chemistry*, *285*(15), 11178–11187.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., … Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* , *26*(18), 2336–2337.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

Pushkin, A., Abuladze, N., Lee, I., Newman, D., Hwang, J., & Kurtz, I. (1999). Mapping of the human NBC3 (SLC4A7) gene to chromosome 3p22. *Genomics*, *57*(2), 321–322.

Riebeling, C., Allegood, J. C., Wang, E., Merrill, A. H., Jr, & Futerman, A. H. (2003).

Two mammalian longevity assurance gene (LAG1) family members, trh1 and trh4, regulate dihydroceramide synthesis using different fatty acyl-CoA donors. *The Journal of Biological Chemistry*, *278*(44), 43452–43459.

Scerri, T. S., Quaglieri, A., Cai, C., Zernant, J., Matsunami, N., Baird, L., … Bahlo, M. (2017). Genome-wide analyses identify common variants associated with macular telangiectasia type 2. *Nature Genetics*. https://doi.org/10.1038/ng.3799

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., … Soranzo, N. (2014). An atlas of genetic influences on human blood metabolites. *Nature Genetics*, *46*(6), 543–550.

Sim, X., Jensen, R. A., Ikram, M. K., Cotch, M. F., Li, X., MacGregor, S., … Wong, T. Y. (2013). Genetic loci for retinal arteriolar microcirculation. *PloS One*, *8*(6), e65804.

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., … Social Science Genetic Association Consortium. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, *50*(2), 229–237.

Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, *8*(1), 1826.

Wilson, E. R., Kugathasan, U., Abramov, A. Y., Clark, A. J., Bennett, D. L. H., Reilly, M. M., … Kalmar, B. (2018). Hereditary sensory neuropathy type 1-associated deoxysphingolipids cause neurotoxicity, acute calcium handling abnormalities and mitochondrial dysfunction in vitro. *Neurobiology of Disease*, *117*, 1–14.

Xie, W., Wood, A. R., Lyssenko, V., Weedon, M. N., Knowles, J. W., Alkayyali, S., … Walker, M. (2013). Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. *Diabetes*, *62*(6), 2141–2150.

Zitomer, N. C., Mitchell, T., Voss, K. A., Bondy, G. S., Pruett, S. T., Garnier-Amblard, E. C., … Riley, R. T. (2009). Ceramide synthase inhibition by fumonisin B1 causes accumulation of 1-deoxysphinganine: a novel category of bioactive 1-deoxysphingoid bases and 1-deoxydihydroceramides biosynthesized by mammalian cell lines and

animals. *The Journal of Biological Chemistry*, *284*(8), 4786–4795.

Zuellig, R. A., Hornemann, T., Othman, A., Hehl, A. B., Bode, H., Güntert, T., … Sonda, S. (2014). Deoxysphingolipids, novel biomarkers for type 2 diabetes, are cytotoxic for insulin-producing cells. *Diabetes*, *63*(4), 1326–1339.

Zuellig, R.A. et al., 2014. Deoxysphingolipids, novel biomarkers for type 2 diabetes, are cytotoxic for insulin-producing cells. *Diabetes*, 63(4), pp.1326–1339.

3.3 Discussion

The work presented above highlighted how serine depletion is likely to be causative for disease and progression. The updated schematic of the study findings is presented in **Figure 34**. Using the framework of Mendelian Randomization we found that genetic influences known to affect serine abundance were associated with MacTel risk. Additionally, serine was highlighted as a potential driver of disease progression. Using the same strategy, we were also able to discard the hypothesis of a causal involvement of glycine/threonine depletion or retinal vasculature calibre on MacTel aetiology and prioritise a possible role of alanine abundance on disease risk. We also confirmed that MacTel heterogeneity is likely to be driven by several genetic influences, some of which are not related to serine abundance or more generally to metabolic traits. By exploiting modern post-GWAS techniques we were able to additionally identify new risk loci for the disease, confirm the involvement previously identified loci and partition them into functional groups. Lastly, the work highlighted a possible inverse causal relationship between MacTel and its comorbidity trait type 2 diabetes and found that the latter might be caused by MacTel, rather than vice-versa.

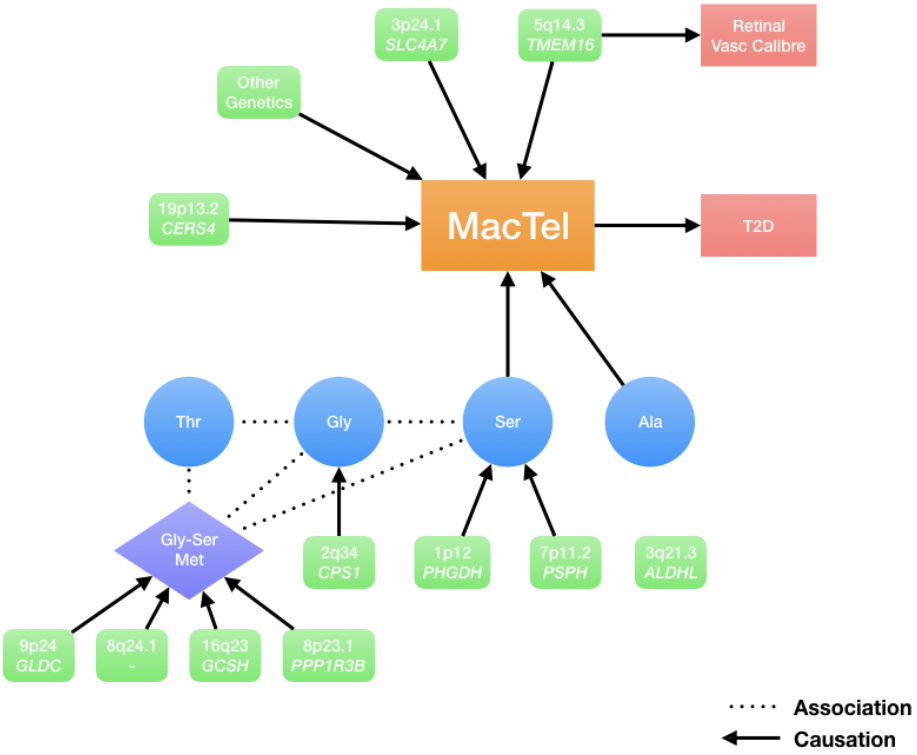


Figure 30: Main findings schematics displaying the drivers and traits associated with MacTel. Genetic traits are displayed as green rounded squares. Metabolites are displayed as blue circles. Metabolic pathways are displayed as purple diamonds. The phenotypic traits are displayed as red squares. Associations are represented by dotted black lines. Causality is indicated by unidirectional solid black arrows.

4 Deep investigation of MacTel metabolic signatures through untargeted metabolomics

4.1 Introduction and study aims

The results presented in Chapters 2 and 3 provide substantive evidence of metabolic disturbances occurring in MacTel patients. Chapter 2 focused only on a few metabolites while Chapter 3 explored metabolic signature only through their genetic components. The evidence presented so far suggests that depletion of serine is likely to have a central role for MacTel occurrence, hence it is important to explore what other metabolic signatures might be present in patients affected by this retinal disease. The following chapter presents the first untargeted metabolomics study performed on MacTel disease. This was a deep investigation of serum metabolic features present in MacTel patients using statistical approaches to uncover potential disease biomarkers as well as elucidating metabolic mechanisms and pathways that could be targeted for therapeutic interventions. This is presented below as a paper under review with Scientific Reports.

4.2 Extract from Bonelli et al, Systemic lipid dysregulation is a novel risk factor for macular neurodegenerative disease

Note: The following manuscript contains figures which are indexed differently from the rest of the thesis as this subchapter is presented in the same formatting as the submitted manuscript.

Systemic lipid dysregulation is a novel risk factor for macular neurodegenerative disease.

Authors

Roberto Bonelli [a,b], Sasha M Woods [c], Brendan R. E. Ansell [a,b], Tjebo FC Heeren [c,d], Catherine A. Egan [d], Kamron N. Khan [e], Robyn Guymer [f], Jennifer Trombley [g], Martin Friedlander [g,h], Melanie Bahlo [a,b], Marcus Fruttiger [c*].
a The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, 3052, Victoria, Australia
b Department of Medical Biology, The University of Melbourne, Melbourne 3010, Victoria, Australia
c UCL Institute of Ophthalmology, University College London, 11-43 Bath St, London EC1V 9EL, United Kingdom
d Moorfields Eye Hospital NHS Foundation Trust, City Road, London EC1, United Kingdom
e The Leeds Teaching Hospitals NHS Trust, St. James's Hospital, Leeds, LS9 7TF, United Kingdom
f Center for Eye Research Australia, Royal Victorian Eye and Ear Hospital, and Ophthalmology, Department of Surgery, 32 Gisborne St, East Melbourne VIC 3002, Australia
g Lowy Medical Research Institute, La Jolla, California, USA
h The Scripps Research Institute, La Jolla, California, USA

* Corresponding author: Marcus Fruttiger,
Address: UCL Institute of Ophthalmology, University College London, 11-43 Bath St, London EC1V 9EL, United Kingdom.
Phone: 0044 20 7608 6872
Email: m.fruttiger@ucl.ac.uk

ORCID Identifiers:

Roberto Bonelli: https://orcid.org/0000-0003-2676-1230
Sasha Woods: https://orcid.org/0000-0001-5101-3053
Brendan R. E. Ansell: https://orcid.org/0000-0003-0297-897X
Tjebo FC Heeren: https://orcid.org/0000-0001-5297-2301
Catherine A. Egan: https://orcid.org/0000-0001-5593-1169
Kamron N. Khan: https://orcid.org/0000-0001-8654-1323
Robyn Guymer: https://orcid.org/0000-0002-9441-4356
Martin Friedlander: https://orcid.org/0000-0003-4238-9651
Melanie Bahlo: https://orcid.org/0000-0001-5132-0774
Marcus Fruttiger: https://orcid.org/0000-0002-6962-5485

Abstract

Background: Macular Telangiectasia type 2 (MacTel) is an uncommon bilateral retinal disease, in which glial cell and photoreceptor degeneration leads to central vision loss. The causative disease mechanism is largely unknown, and no treatment is currently available. A

previous study found variants in genes associated with glycine-serine metabolism (*PSPH, PHGDH* and *CPS1*) to be associated with MacTel, and showed low levels of glycine and serine in the serum of MacTel patients.
Methods: We used a global metabolomics platform in a case-control study to comprehensively profile serum from 60 MacTel patients and 58 controls.
Results: Analysis of the data, using innovative computational approaches, revealed a detailed, disease-associated metabolic profile with broad changes in multiple metabolic pathways. This included alterations in the levels of several metabolites that are directly or indirectly linked to glycine-serine metabolism, further validating our previous genetic findings. However, we also found changes unrelated to PSPH, PHGDH and CPS1 activity. Most pronounced, levels of several lipid groups were altered, with increased phosphatidylethanolamines being the most affected lipid group. Additionally, assessing correlations between different metabolites across our samples revealed putative functional connections. Correlations between phosphatidylethanolamines and sphingomyelin, and glycine-serine and sphingomyelin, observed in controls, were reduced in MacTel patients, suggesting metabolic re-wiring of sphingomyelin metabolism in MacTel patients.
Conclusions: Our findings provide novel insights into metabolic changes associated with MacTel and implicate altered lipid metabolism as a novel contributor to this retinal neurodegenerative disease.

Keywords:

Metabolomics | Macular Telangiectasia Type 2 | Serum metabolites | Lipids | Serine | Glycine

Introduction

Macular telangiectasia type 2 (MacTel) is an uncommon, bilateral neurodegenerative retinal disease affecting between 0.004% and 0.1% of the population [1,2]. It is characterized by alterations of the macular capillary network and neurosensory atrophy beginning temporal to the fovea, eventually affecting the so-called “MacTel area”; an oval area approximately 3mm across the temporal-nasal axis and 2mm across the superior-inferior axis centred on the fovea and of similar size in all patients [3]. Symptoms typically start in the fifth and sixth decade of life, most commonly with reading difficulties and distortions [3–6]. The pathogenic mechanism of this disease is still not fully understood, but post-mortem histopathological studies show abnormalities in the retinal pigment epithelium (RPE) throughout the retina [7] and a complete loss of Müller cells specifically in the MacTel area; which also contains some regions of photoreceptor loss [8]. A recent phase 2 clinical trial with an ocular implant secreting ciliary neurotrophic factor showed promising results in delaying disease progression, but no evidence of recovery [9]. No other therapeutic treatments are currently available.

Several factors suggest that MacTel has a substantial genetic component. The disease occurs bilaterally and is heritable based on studies of monozygotic twins, siblings and families [3,10–13]. We recently performed a Genome-Wide Association Analysis (GWAS) for MacTel, which identified five genetic loci associated with the disease [14]; of which four are associated with glycine and serine abundance in serum [15–17]. Furthermore, we observed lower glycine and serine levels in the serum of MacTel patients relative to controls [14]. However, glycine and serine play a role in several different metabolic pathways, and thus the MacTel disease mechanism remains unclear; in particular, which factors contribute to the local retinal specificity of the disease in otherwise healthy individuals. The degree to which metabolites, other than glycine and serine, may contribute to MacTel also requires further investigation.

To comprehensively characterise the metabolomic profile of MacTel patients, we analysed hundreds of serum metabolites using a global metabolomics platform. The vast data captured by this technology requires advanced data processing and statistical analysis [18,19]. Here we employed state-of-the-art statistical methodologies [18,20,21], to characterise the impact of reduced glycine and serine in a broad metabolic context, and identify novel metabolic pathways associated with MacTel.

Materials and Methods

Study participants and serum collection

All experiments were conducted according to the principles expressed in the Declaration of Helsinki. All participants gave informed consent and the study was approved by the Research Ethics Office Bromley, UK (study number 05/Q0504/101). Blood serum was collected from 60 random patients from the MacTel Natural History and Observation Study (NHOS). Control samples were collected from 58 unrelated individuals who do not suffer from MacTel. We attempted to roughly match the control and MacTel patient cohorts for age, gender, diabetic

status and ethnicity. However, this match was not perfect (**Table S1**), and differences were corrected for in the downstream statistical analysis.

The number of samples in our study was based on previous experience by Metabolon (Durham, USA) regarding effect size of serum metabolomic analysis on the analytical platform used. Samples were collected at different clinical sites (Moorfields Eye Hospital London, UK; Royal Victorian Eye and Ear Hospital and Ophthalmology, Melbourne, Australia; Scripps Health Facility, Scripps Clinic Torrey Pines, La Jolla, USA, and St. James' Hospital, Leeds, UK) according to a standardised protocol. All individuals fasted overnight, and blood was taken before noon. Around 5ml of blood was collected in a clot activating vacutainer tube (Vacutainer Plastic SST II Advance Tube with Gold Hemogard Closure, Becton Dickinson), left at room temperature for 30min and then centrifuged for 5 min at 1200g. The supernatant was collected, frozen and stored at -80°C.

Metabolite measurements

Serum metabolites were measured by Metabolon (Durham, USA). Briefly, this involved initial protein precipitation with methanol under vigorous shaking for 2 min (Glen Mills GenoGrinder 2000) followed by centrifugation. The resulting extract was divided into five fractions: two for analysis by two separate reverse phase (RP)/UPLC-MS/MS methods with positive ion mode electrospray ionization (ESI); one for analysis by RP/UPLC-MS/MS with negative ion mode ESI; and one for analysis by HILIC/UPLC-MS/MS with negative ion mode ESI. Raw data was extracted, peak-identified and QC processed using Metabolon's hardware and software. Compounds were identified by comparison to library entries of purified standards and peaks were quantified using the area-under-the-curve technique, providing relative abundances of 946 individual metabolites.

Metabolomics Data Processing

Missing values for some metabolites indicated levels below the detection limit of the Metabolon platform. For this reason, missing abundances were imputed with the minimum value for each metabolite following Metabolon's standard imputation protocol. In subsequent analyses, we discovered how imputed minimum values for each metabolite were driving a substantial amount of variance captured by principal components (PCs). Because of this, we decided to discard 194 metabolites which had more than 20% of their total values missing from the analysis; using the previously proposed "80% rule" [22,23]. Metabolomics missingness in controls correlated with missingness in patients, which ensured compatibility between patients and control (**Fig. S2**) and ensured that missingness would not confound the disease signal. The average missingness per subject was 13.3% with SD 1.5%. Boxplot outlier detection was performed on the distribution of missing values among subjects: this analysis did not detect any outlier subjects for missingness, and no particular subject was excluded for excessive amount of missingness.

A total of 738 metabolites passed quality control steps and were further analysed. We used the R software package limma [24] for further normalisation of the data, not captured by the initial area-under-the-curve normalization. This software package was initially developed for gene expression analysis but has been cross-purposed for metabolite analyses [24–26]. To

interrogate whether the heterogeneity in the remaining metabolites was driven by the diversity between subject demographics we visually inspected the first two principal components against all available covariates. Visual exploration of principal components plots ensured that no unwanted variation was observable after the scaling and normalisation step (**Fig. S3**). Lastly, we scaled each metabolite to have zero mean and unit standard deviation to ensure comparability for each specific metabolomic result.

Statistical Analysis

Differences in demographics between patients and controls were tested using Chi-squared tests (for dichotomous demographics) and t-tests with Welch correction for unequal variances (for continuous variables). We tested each metabolite against disease status using the limma package for the statistical software, R [27]. This package has been widely used in gene expression study and is based on an empirical Bayes approach; whose properties for the analysis of both microarray data and RNAseq data have been described elsewhere [27,28]. We corrected each p-value for false discovery rate by applying the Benjamini-Hochberg p-value correction procedure on the nominal uncorrected p-values using the R function p.adjust [29]. All metabolites with a corrected p-value less than the false discovery rate cut off of 0.05 were considered as significant. Each model corrected for all available covariates which included sex at birth, age, diabetes status, ethnicity and BMI. A clustered heatmap of significant metabolites is presented in (**Fig. S4**).

After defining the groups, we tested for enrichment by using a popular tool for gene enrichment pathway analysis - ROAST - available in the limma package. The properties and details of this tool have been discussed elsewhere [30]. We corrected for false discovery rate by correcting each p-value using the Benjamini-Hochberg procedure. All groups with a corrected p-value of less than 0.05 were considered as significant. To account for groups in which metabolites were changed in opposite directions, we represented the abundance of grouped metabolites as a single value (the first principal component) and performed differential expression analysis using the same method applied to individual metabolites.

To test for differential co-abundance, we firstly calculated the correlation between pairs of metabolites within patients and within controls. However, metabolomics correlation can be confounded by a set of external factors. To account for this, we firstly residualised all metabolomics profiles for each covariate jointly. This was achieved by regressing each metabolite on the set of covariates and MacTel status using a linear regression model and sequentially extracting the model residuals. This approach ensured that no correlation between metabolite was either created or masked by the covariates. After residualisation, we excluded from the analysis all those metabolites pairs for which the absolute value of their correlation was less than 0.5. We then tested for differential co-abundance between pairs of metabolites similarly to previous work [31] by using the Fisher R to Z transformation as defined in the Psych package [32]. We corrected for false discovery rate by correcting each retained p-value using Benjamini-Hochberg procedure. Given the very high number of correlation pairs tested and the reduced sample size, all pairs with a corrected p-value less than 0.1 were considered as significantly differentially co-abundant. Following the definition of previous work on gene expression by Jiang et al. [31] we explored the hypothesis that specific groups might present an abnormal amount of differential co-abundance. Accordingly, we discarded all

metabolites belonging to the xenobiotic super classification from this analysis. To test the hypothesis of groups enriched with differential co-abundance, we used a binomial test. The number of successes was the number of significant differential co-abundant pairs which contained a metabolite in that group. The number of tests was the total number of co-abundant pairs tested which contained a metabolite belonging to that group. Lastly, the probability parameter was defined as the ratio between all significantly differentially co-abundant pairs in the dataset - 161 - and all tested pairs 19.127 (p_parameter= 0.008). We corrected for false discovery rate by correcting each p-value using Benjamini-Hochberg procedure. All metabolites with a corrected p-value less than 0.05 were considered as enriched.

Results

Metabolite levels

We conducted an untargeted global metabolomic analysis (Metabolon, Inc.) of serum from 60 extensively phenotyped MacTel patients and 58 healthy controls, measuring the relative levels of a total of 946 known metabolites, of which 738 survived quality control. The patient and control cohorts were roughly matched for demographics (**Table S1**), and statistical tests (see Materials and Methods) revealed no significant differences between them, regarding average age (p=0.06), gender (p=0.14), body mass index (BMI, p=0.34) and - since diabetes is a comorbidity of MacTel [33,34] - diabetic status (p=0.22). We were not able to completely match ethnicity status between cases and controls (p=0.009). Nevertheless, we used normalisation and multivariate regression strategies to correct for all available covariates (age, gender, diabetes status, ethnicity, BMI and collection site, as described in Materials and Methods). The fully normalised dataset (**Table S2**) was then analysed for differential abundance of individual metabolites in patients versus controls. This revealed 49 metabolites with significantly lower serum concentrations in MacTel patients compared to controls, and 72 with elevated serum concentrations (p<0.05, corrected for false discovery rate, FDR) (**Fig. 1**).

Glycine and serine were the first and third most depleted metabolites in MacTel patients (log(2) fold changes (logFC) of -1.31 and -1.03, respectively). This agrees with our recent GWAS [14], which also showed reduced glycine/serine concentrations in MacTel patients, using a subset of the samples presented here. The second and fourth most changed metabolites were gamma-glutamylglycine and alpha-ketoglutarate (logFC of -1.08 and -0.98, respectively), both of which are linked to glycine-serine metabolism (see discussion). However, we also found many changed metabolites belonging to other metabolism groups (**Fig. 1, Fig. 3**). For instance, we found reduced levels of arginine (logFC=-0.76), ornithine (logFC=-0.59) and guanidinoacetate (logFC=-0.47), which - together with glycine - are needed for creatine biosynthesis (see discussion). Similarly, methionine (logFC=-0.58) and betaine (logFC=-0.53), which are linked to cysteine-methionine metabolism (see discussion), were reduced in MacTel patients. In contrast, the majority of measured lipids were increased. Furthermore, of the 121 significantly changed metabolites, 88 were lipids (73%), whilst the total dataset of 738 metabolites contained 45% lipids, suggesting a disproportional impact on lipid metabolism in MacTel.

To fully assess whether the observed changes affected specific metabolic pathways, we divided all metabolites into 50 functional groups -largely reflecting pathways assigned by the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database - and tested (described in Materials and Methods) whether any groups presented enrichment of differential abundance (**Fig. 2, Table S3**).

The most differentially abundant metabolite group in MacTel patients was glycine-serine-threonine metabolism (p=2.5E-5), with 7 of 11 measured metabolites depleted. The second most significantly different group was the phosphatidylethanolamines (p=8.8E-5), with 13 out of 14 metabolites upregulated in MacTel patients. Other differentially abundant groups included long chain fatty acids (increased, p=0.02) and diacylglycerols (increased, p=0.012), as well as changes in alanine-asparagine (reduced, p=0.012), methionine-cysteine (mixed, PC p=0.018) and benzoate metabolism (reduced p=0.002). We further detected several pronounced lipid group differences including increases in lysophosphatidylethanolamines (p=0.0054), diacylglycerols (p=0.012), monoacylglycerols (p=0.463) and long chain fatty acids (p=0.020) (**Table S3**). Although the majority of lipids were increased in MacTel patients, etherlipids with a choline headgroup were markedly reduced (p=0.020). A second group of lipids showing a reverse trend of general lipid increases in MacTel patients were the sphingomyelins (p=0.012), where 20 of the 21 measured species were reduced (7 of which were significant; **Fig. 1**).

Correlations between metabolites

Having established that the abundance of several metabolites was altered in MacTel patients we investigated how metabolites correlated with each other (co-abundance) across patients and controls. This technique can reveal molecular interactions that change in the context of disease [35] and can inform potentially dysregulated biochemical mechanisms [31,36,37]. Differential correlation testing was limited to those metabolite pairs which were significantly correlated after correction for multiple comparisons in either patients or controls. Whereas most correlations between metabolites were similar across both groups (**Fig. S1**), three were significantly different in MacTel patients compared to the controls (**Table S4**). Orotate and orotidine were correlated in patients (r=0.73) but not controls (r=0.02; p=0.001); and the correlation between xanthine and orotidine was negative in controls (r=-0.43) but positive in patients (r=0.43; p=0.001). These metabolites all function within pyrimidine metabolism. However, closer inspection of these results revealed a potential xenobiotic confound, driven by four patients with elevated oxypurinol levels, likely due to Allopurinol drug exposure, which is known to disturb pyrimidine metabolism.

We additionally tested for enrichment in differential co-abundance at the metabolite group (pathway) level. Tests of enrichment for differential co-expression were performed using a binomial framework [31]. Of 45 metabolite groups, seven were enriched in differentially correlated metabolite groups in patients compared to controls (p<0.05; **Fig. 4, Table S5**). Among these, methionine-cysteine metabolism and the sphingomyelin group overlapped with our metabolite abundance results. The sphingomyelin group exhibited the most homogeneous pattern of differential co-abundance, characterised by a unique enrichment in disrupted connections. Specifically, differential correlation between sphingomyelins and the most

differentially abundant groups - phosphatidylethanolamines and glycine-serine-threonine pathway metabolites - were driving this result. Positive correlations between the sphingomyelins and serine-glycine pathway metabolites, observed in controls, were lost in MacTel patients; while negative correlations between sphingomyelins and the phosphatidylethanolamines, were reduced or lost in patients (**Fig. 5**). This striking result represents disruption of the normal metabolic links between serine, phosphatidylethanolamines and sphingomyelins in sphingolipid metabolism of MacTel patients.

Discussion

Data analysis tools

Isolating relevant signals in metabolomics data requires state-of-the-art statistical methods. For this study we used limma software [27], the gold standard in gene expression analysis, to normalize for confounding factors, construct sample quality weights and perform multivariate modelling. Additionally, by using an empirical Bayesian framework that "borrows" information between metabolites, the software ensures maximal discovery power even for extremely variable metabolites [27,28]. To extend our finding beyond simple biomarkers, we performed enrichment analysis using Roast [30] which takes into account metabolite abundances rather than summary statistics [28,30]. In addition, we conducted semi-supervised principal components analysis to assess significance in groups containing metabolites with mixed correlations - which might have been missed by Roast - as well as exploring metabolic network changes through differential co-abundance analysis. Although the computational tools we employed here have so far not been commonly used for metabolomics data analysis, in this study we demonstrate that their deployment in the field of metabolomics can be extremely useful and powerful.

Genetic variants in PSPH, PHGDH and CPS1 deeply impact metabolic profiles in MacTel patients

Our previous GWAS [14] identified MacTel disease risk-associated single nucleotide polymorphisms (SNPs) within three genes, *PSPH* (encoding phosphoserine phosphatase), *PHGDH* (encoding phosphoglycerate dehydrogenase) and *CPS1* (encoding carbamoyl-phosphate synthase), which are all known to contribute to glycine-serine metabolism. PSPH and PHGDH catalyse consecutive steps in the serine biosynthesis pathway (**Fig. 3**), and mutations in *PSPH* are associated with reduced plasma serine levels [38]. Similarly, SNP rs477992 reduces *PHGDH* transcription [39] and serine plasma levels [16,17]. This matches well with the clear reductions in glycine and serine in MacTel patient serum we are presenting here. Furthermore, PHGDH is also known to convert alpha-ketoglutarate to 2-hydroxyglutarate [40,41]. While the former was only marginally increased in patient serum (logFC=0.40), the latter was strongly depleted (logFC=-0.98), further supporting a likely PHGDH defect in MacTel patients. Serine depletion, caused by dysfunctional PSPH and PHGDH, increases conversion of glycine to serine (**Fig. 3**) [42], explaining the reduced glycine levels, at least in part.

Another likely mechanism contributing to low glycine in MacTel patients is based on the activity of CPS1, which is connected to the urea cycle and creatine biosynthesis (**Fig. 3**). CPS1 converts ammonia and bicarbonate into carbamoyl phosphate, which then feeds the urea cycle. The C allele of SNP rs715 (located in the 3' UTR of *CPS1*; estimated MAF = 0.24) has been found in a previous independent study to be strongly associated with increased glycine serum levels [16]. *CPS1* is also implicated, via GWAS studies, in modulating creatine [16], arginine and ornithine levels [43]. These effects are based on the role of CPS1 in creatine biosynthesis, which requires urea metabolites. In a key intermediary step, glycine and arginine are converted to guanidinoacetate and ornithine (**Fig. 3**), all four of which we found to be reduced in MacTel serum (logFC=-1.31, logFC=-0.76, logFC=-0.46 and logFC=-0.59, respectively). Reduced CPS1 activity slows the urea cycle and the linked guanidino-acetate production. This likely results in reduced glycine consumption and may explain the protective role of rs715(C) in MacTel [14].

Mutations in *CPS1* might also explain the depleted threonine levels (logFC=-0.96) in our MacTel patients (**Fig. 2 and 3**) since a previous GWAS has linked *CPS1* with threonine plasma levels [43]. However, the biochemical connections between CPS1 and threonine are not clear. Whilst most mammals can directly convert glycine to threonine, the enzyme responsible for this reaction (threonine aldolase) has lost function in humans [44], making threonine an essential amino acid. Alternatively, it is conceivable that microbiome influences may contribute to glycine-threonine conversion [45].

In addition, several further metabolites linked to *CPS1* via GWAS [43,46] were reduced in our MacTel samples (asparagine, logFC=-0.85; glutamine, logFC=-0.74 and betaine, logFC=-0.53). Asparagine and glutamine have been linked to CPS1 by a GWAS [43], matching the correlations found in our study (glycine-asparagine, r=0.65 and glycine-glutamine, r=0.51, in controls). As glutamine is needed to create intracellular asparagine, which in turn is needed for serine uptake [47], depletion of these metabolites - as observed in our study - is likely to additionally contribute to the low serine availability observed in MacTel. However, the correlation between these metabolites was not changed in MacTel patients compared to controls (glycine-asparagine, r=0.63 and glycine-glutamine, r=0.56). Of interest, there was a trend towards reduced correlation between glycine and betaine (r=0.56 in controls, r=0.16 in MacTel patients), but this change did not attain statistical significance (p=0.22).

In addition, we observed a trend towards changed correlations between serine and pyruvate, which rose from r=-0.16 in controls to r=0.39 in patients. Although the change did not reach statistical significance (p=0.13), it is interesting that these two metabolites are linked via serine dehydratase (SDS), which converts serine to pyruvate (**Fig. 3**). An increased correlation between serine and pyruvate possibly indicates a more pronounced usage of this pathway in MacTel patients, which would reduce serine levels. Of note, SDS can also degrade threonine to α-ketobutyrate and ammonia (**Fig. 3**), which might be reflected by the increased α-ketobutyrate levels in MacTel serum due to a potential increase of SDS activity in our patients.

Changes in metabolites related to cysteine/methionine metabolism imply oxidative stress and phospholipid species bias in MacTel

GWAS has also linked *CPS1* to the metabolites betaine, choline and homocysteine [46,48], which are all relevant for cysteine/methionine metabolism (**Fig. 3**). In MacTel patients, betaine and choline were both reduced (logFC=-0.53, logFC=-0.46, respectively), which might relate to the fact that methionine was also lower in MacTel patients (logFC=-0.58). The strong correlations we observed between methionine and asparagine (r=0.73 in controls, r=0.69 in MacTel patients) and between methionine and threonine (r=0.61 in controls, r=0.74 in MacTel patients) support the notion of a potential involvement of CPS1. However, the mechanism by which CPS1 could influence methionine levels is not known.

Methionine is an essential amino acid and can be recycled from homocysteine via two different pathways. One depends on betaine (as mentioned above, low in MacTel patients), while the other requires 5,10-methylenetetrahydrofolate (CH2-THF in **Fig. 3**), which was not directly measured in our study, but is likely to be reduced given its close metabolic links to serine and glycine [49]. Additionally, low choline and betaine observed in MacTel patients increases the reliance of the methionine cycle on one-carbon metabolism, adding a further demand on glycine. Furthermore, histidine - also used to add one carbon to tetrahydrofolate - was also reduced in MacTel patients (logFC=-0.55). Together, these findings suggest lower methionine cycle capacity in MacTel patients compared to controls.

In this context it is interesting that MacTel patients exhibited increased concentrations of alpha-ketobutyrate (logFC=0.59) and 2-hydroxybutyrate (logFC=0.63), which may relate either to increased threonine degradation or to increased glutathione production (**Fig. 3**). Possible reasons for the latter are increased oxidative stress or detoxification in the liver [50]. Increased glutathione synthesis consumes serine and glycine. Furthermore, it also limits the supply of cysteine, diverting homocysteine away from the transmethylation pathway towards glutathione synthesis, leading to a stressed transmethylation pathway.

A key product of the methionine cycle is S-adenosylmethionine (SAM), required for transmethylation reactions. The conversion of phosphatidylethanolamine to phosphatidylcholine requires three SAM molecules and is, therefore, particularly affected by methionine cycle limitations. This strongly agrees with the aforementioned increase of phosphatidylethanolamines found in MacTel patients. Although phosphatidylcholine levels were nominally elevated (p=0.076), the phosphatidylethanolamine increase was much more pronounced (p=8.8e-05), apparent in nearly all measured phosphatidylethanolamine species (**Fig. 3, Table S3**). Interestingly, ether lipids with an ethanolamine head group were less severely reduced, mirroring the shift in the ethanolamine/choline ratio mentioned here.

The lower abundance of methionine observed in MacTel patients may be connected to changes in both glutathione and phosphatidylcholine synthesis. As a relative lack of methionine may impair glutathione synthesis, this might expose MacTel patients to a higher oxidative stress load. Additionally, the methionine derivative SAM provides the methyl group substrates required to produce choline from ethanolamine. In fact, we found increased phosphatidylethanolamine in MacTel patients, suggesting an imbalance in the

phosphatidylethanolamine/phosphatidylcholine ratio. Further, substantial reduction in choline relative to ethanolamine etherphosphoilipids in patients is also consistent with methionine depletion.

Lipid dysregulation is a novel disease risk factor for MacTel

Several lipid groups were significantly changed in MacTel patients; with increased phosphatidylethanolamines, lysophosphatidylethanolamines and diacylglycerols, and decreased sphingomyelins (**Table S3**). These changes cannot be linked (directly or indirectly) to the activity of PSPH, PHGDH and CPS1, based on current understanding of human metabolic pathways. It is, therefore, plausible that the observed lipid dysregulation in MacTel patients represents a novel MacTel risk factor, in addition to the previously identified risk factors related to glycine-serine metabolism. The abnormal lipid levels in MacTel patients may be caused by as yet unidentified genetic risks, as well as by environmental or dietary influences.

Sphingomyelins link dysregulated lipids and serine metabolism

Differential co-abundance analysis identified sphingomyelins as an important metabolic group for MacTel. Not only were sphingomyelins depleted in MacTel, but metabolic connections - indicated by correlations of metabolites - between this group with the glycine-serine metabolism group, and the phosphatidylethanolamine group, were lost (**Fig. 5**). Although serine was depleted in MacTel patients, phosphatidylethanolamines and their fatty acid constituents were enriched. This striking change in correlation is particularly interesting in context of our recent finding that sphingolipid metabolism is an important component in MacTel [51]. Serine is an essential building block for the biosynthesis of sphinganine, which forms the backbone of all sphingolipids [52]. We found that low serine in the serum of MacTel patients correlates with increased levels of deoxysphingolipids. These atypical sphingolipids are a risk factor for MacTel [51] and are known to have cytotoxic properties [52]. How exactly they contribute to retinal pathology is not fully understood yet, but it is well established that dysregulation of sphingolipid metabolites can differentially regulate apoptosis and autophagy [53]. In this context it is also interesting that ceramides were partially enriched in MacTel patients (2 out of 6 measured). Ceramides can act as pro-apoptotic signalling molecules, and it has been shown that blocking ceramide biosynthesis prevents photoreceptor cell death in a mouse model of retinitis pigmentosa [54].

MacTel metabolic dysregulation might affect type 2 diabetes risk

Previous studies have observed a high prevalence of type 2 diabetes among MacTel patients [34,55]. However, the mechanism explaining this association is not understood. We therefore roughly matched for diabetic status in our MacTel and control cohorts (**Table S1**), and used a computational approach (described in Materials and Methods) to correct for any remaining imbalances. Despite this, our study identified an intriguing overlap between MacTel metabolites and previously described metabolic phenotypes associated with type 2 diabetes [56–58]. For instance, MacTel patients displayed increased fatty acids, diacylglycerols and phosphatidylethanolamines, as well as reduced etherlipids, glycine and glutamine, a similar

pattern to that observed in type 2 diabetes metabolomics studies, hinting at a potential mechanistic link between MacTel and diabetes. One possible such interaction might be mediated via the formation of deoxysphingolipids mentioned above, which apart from MacTel have also been linked to type 2 diabetes mellitus [59–61].

Using serum metabolomics to study a complex eye disease

MacTel is a genetically complex and moderately clinically heterogeneous disease involving focal degeneration of retinal glia and photoreceptors, and vascular damage of the central retina (macula). MacTel heterogeneity may arise from a complex mechanism impacting multiple biological pathways. Our previous study [14] provided initial evidence that the disease was associated with global alterations of metabolism, determined in part by complex genetic contributions. To explore this hypothesis, we comprehensively compared the metabolome of MacTel patients to that of healthy controls. Metabolomics is an emerging field and has been recognised as a powerful tool in ophthalmological research [62,63]. In this study, we used serum samples, although MacTel is a retinal disease not characterised by any known systemic pathology. This may appear counterintuitive, as retinal metabolism is considered to be isolated from the peripheral blood circulation due to the blood-retinal barrier. However, MacTel is associated with systemic hyperglycaemia [34,55], and metabolomics profiles of blood samples have been used to study other retinal diseases, identifying changes as potential metabolic biomarkers or interrogating broad dysregulations [64–67]. Furthermore, serum samples are more readily available than retina samples in people with the disease, and metabolomics platforms for analysis are readily and commercially available. Lastly, a recent metabolomics study performed on mouse models demonstrate that systemic serine and glycine changes mirroring MacTel metabolic dysregulation reflect deeply in mice retina [50].

It is important to keep in mind that, although our study has identified systemic risk factors for the disease, the metabolic changes in MacTel retina are so far unknown. It is also unknown by what route such changes might affect the retina. It is possible that the underlying genotypes that cause the systemic changes - for example due to altered liver metabolism – affect the retina indirectly via the circulation. On the other hand, it is also possible that the genes that contribute to the systemic changes may have independent functions in the retina where they directly contribute to the disease.

Conclusion

In summary, in this study we have revealed putative functional relationships between multiple metabolite groups and MacTel disease using state-of-the-art analyses. Several of the changes we observed in this study confirm our previously identified genetic MacTel findings where we implicated *PSPH, PHGDH and CPS1*; however, many other changes are novel and represent novel risk factors for this disease. Our study provides not only a foundation for future genetic and experimental analyses of MacTel pathobiology, but also serves as a template for the use of computational approaches to global metabolomics data to investigate diseases with complex aetiologies.

Declarations

Ethics approval and consent to participate

All experiments were conducted according to the principles expressed in the Declaration of Helsinki. All participants gave informed consent and the study was approved by the Research Ethics Office Bromley, UK (study number 05/Q0504/101).

Availability of data and material

Raw data not included in this published article will be included as supplementary information.

Competing interests

The authors declare no conflict of interest.

Funding

Support for this study came from the Lowy Medical Research Institute, National Institute of Health Research (NIHR) Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom (partial funding, CE and MF), and National Institute for Health Research (NIHR). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. This work was also made possible through the Victorian State Government Operational Infrastructure Support and Australian Government National Health and Medical Research Council (NHMRC) independent research Institute Infrastructure Support Scheme (IRIISS). RB was supported by the Melbourne International Research Scholarship. BREA was supported by an NHMRC early career Fellowship (1157776). RB and TFCH were supported the Lowy Medical Research fellowship. MB was supported by an NHMRC Senior Research Fellowship (1102971) and Program Grant (1054618). RG was supported by an NHMRC Senior Research Fellowship (1103013). KNK was supported by an NIHR-Rare Disease Fellowship. The funding organizations had no role in the design or conduct of this research.

Authors' contributions

MF(c), CAE, KNK, MB conceived the study, SMW, TFCH, CAE, KNK, RG, JT, MF(h) collected and prepared samples, RB, SMW, BREA, MB, MF(c) analysed data, and all authors contributed to the writing and final approval of the manuscript.

Acknowledgements

We thank the volunteers who participated in this study. We are grateful to Saskia Freytag, Anna Quaglieri, Terry Speed, Gordon Smyth and Mari Gantner for helpful discussion.

References

1. Tham, Y. C. *et al.* Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. *Ophthalmology* **121**, 2081–2090 (2014).
2. Klein, R. *et al.* The Prevalence of Macular Telangiectasia Type 2 in the Beaver Dam Eye Study. *Am. J. Ophthalmol.* **150**, 2705–2710 (2010).
3. Charbel, P. *et al.* Macular telangiectasia type 2. *Prog. Retin. Eye Res.* **34**, 49–77 (2013).
4. Finger, R. P. *et al.* Reading performance is reduced by parafoveal scotomas in patients with macular telangiectasia type 2. *Investig. Ophthalmol. Vis. Sci.* **50**, 1366–1370 (2009).
5. Clemons, T. E. *et al.* The national eye institute visual function questionnaire in the macular telangiectasia (MacTel) project. *Investig. Ophthalmol. Vis. Sci.* **49**, 4340–4346 (2008).
6. Heeren, T. F. C., Holz, F. G. & Issa, P. C. FIRST SYMPTOMS AND THEIR AGE OF ONSET IN MACULAR TELANGIECTASIA TYPE 2. *Retina* **34**, 916–919 (2014).
7. Powner, M. B. *et al.* Fundus-Wide Subretinal and Pigment Epithelial Abnormalities in Macular Telangiectasia Type 2. *Retina* **38**, S105–S113 (2018).
8. Powner, M. B. *et al.* Loss of Müller's cells and photoreceptors in macular telangiectasia type 2. *Ophthalmology* **120**, 2344–2352 (2013).
9. Chew, E. Y. *et al.* Effect of Ciliary Neurotrophic Factor on Retinal Neurodegeneration in Patients with Macular Telangiectasia Type 2: A Randomized Clinical Trial. *Ophthalmology* **126**, 540–549 (2019).
10. Parmalee, N. L. *et al.* Analysis of candidate genes for macular telangiectasia type 2. *Mol. Vis.* **16**, 2718–26 (2010).
11. Parmalee, N. L. *et al.* Identification of a Potential Susceptibility Locus for Macular Telangiectasia Type 2. *PLoS One* **7**, 1–10 (2012).
12. Ronquillo, C. C., Wegner, K., Calvo, C. M. & Bernstein, P. S. Genetic Penetrance of Macular Telangiectasia Type 2. *JAMA Ophthalmol.* **136**, 1158 (2018).
13. Szental, J. A. *et al.* Analysis of glutathione S-transferase Pi isoform (GSTP1) single-nucleotide polymorphisms and macular telangiectasia type 2. *Int. Ophthalmol.* **30**, 645–650 (2010).
14. Scerri, T. S. *et al.* Genome-wide analyses identify common variants associated with macular telangiectasia type 2. *Nat. Genet.* **49**, 559–567 (2017).
15. Xie, W. *et al.* Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. *Diabetes* **62**, 2141–2150 (2013).
16. Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
17. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54–62 (2011).
18. Considine, E. C., Thomas, G., Boulesteix, A. L., Khashan, A. S. & Kenny, L. C. Critical review of reporting of the data analysis step in metabolomics. *Metabolomics* **14**, 7 (2017).
19. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: Beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).
20. Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D. & Lu, L. J. Computational and statistical analysis of metabolomics data. *Metabolomics* **11**, 1492–1513 (2015).
21. Alonso, A., Marsal, S. & Julià, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Front. Bioeng. Biotechnol.* **3**, 1–20 (2015).
22. Bijlsma, S. *et al.* Large-scale human metabolomics studies: A strategy for

data (pre-) processing and validation. *Anal. Chem.* **78**, 567–574 (2006).
23. Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **8**, 1–10 (2018).
24. Frediani, J. K. *et al.* Plasma metabolomics in human pulmonary tuberculosis disease: A pilot study. *PLoS One* **9**, (2014).
25. Fischer, R. *et al.* Discovery of candidate serum proteomic and metabolomic biomarkers in ankylosing spondylitis. *Mol. Cell. Proteomics* **11**, 1–11 (2012).
26. Putluri, N. *et al.* Metabolomic Profiling Reveals Potential Markers and Bioprocesses Altered in Bladder Cancer Progression. *Cancer Res.* **71**, 7376–7386 (2011).
27. Smyth, G. K. limma: Linear Models for Microarray Data. in *, Statistics for Biology and Health* (eds. Gentleman, R. & Carey, V.J., Huber W., Irizarry R.A., D. S.) 397–420 (Springer-Verlag, 2005). doi:10.1007/0-387-29362-0_23
28. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
29. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.* **57**, 289–300 (2007).
30. Wu, D. *et al.* ROAST: Rotation gene set tests for complex microarray experiments. *Bioinformatics* **26**, 2176–2182 (2010).
31. Jiang, Z., Dong, X., Li, Z., He, F. & Zhang, Z. Differential Coexpression Analysis Reveals Extensive Rewiring of Arabidopsis Gene Coexpression in Response to Pseudomonas syringae Infection. *Nat. Publ. Gr.* 1–13 (2016). doi:10.1038/srep35064
32. Revelle, W. An overview of the psych package. (2017). Available at: http://personality-project.org/r/overview.pdf.
33. Clemons, T. E. *et al.* Medical characteristics of patients with macular telangiectasia type 2 (MacTel Type 2) MacTel Project Report No. 3. *Ophthalmic Epidemiol.* **20**, 109–113 (2013).
34. Chew, E. Y., Newsome, D. A. & Fine, S. L. Parafoveal Telangiectasis and Diabetic Retinopathy. *Arch. Ophthalmol.* **104**, 71–75 (1986).
35. Hsu, C., Juan, H. & Huang, H. Functional Analysis and Characterization of Differential Coexpression Networks. *Nat. Publ. Gr.* 1–14 doi:10.1038/srep13295
36. Choi, J. K., Yu, U., Yoo, O. J. & Kim, S. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics* **21**, 4348–4355 (2005).
37. Gov, E. & Arga, K. Y. Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer. *Sci. Rep.* **7**, 1–10 (2017).
38. Veiga-da-Cunha, M. *et al.* Mutations responsible for 3-phosphoserine phosphatase deficiency. *Eur. J. Hum. Genet.* **12**, 163–166 (2004).
39. Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
40. Zhao, G. & Winkler, M. E. A novel alpha-ketoglutarate reductase activity of the serA-encoded 3-phosphoglycerate dehydrogenase of Escherichia coli K-12 and its possible implications for human 2-hydroxyglutaric aciduria. *J. Bacteriol.* **178**, 232–9 (1996).
41. Fan, J. *et al.* Human phosphoglycerate dehydrogenase produces the oncometabolite D-2-hydroxyglutarate. *ACS Chem. Biol.* **10**, 510–516 (2015).
42. Pacold, M. E. *et al.* A PHGDH inhibitor reveals coordination of serine synthesis and one-carbon unit fate. *Nat. Chem. Biol.* **12**, 452–458 (2016).
43. Imaizumi, A. *et al.* Genetic basis for plasma amino acid concentrations based on absolute quantification: a genome-wide association study in the Japanese population. *Eur. J. Hum. Genet.* **27**, 621–630 (2019).
44. Malinovsky, A. V. Reason for indispensability of threonine in humans and other mammals in comparative aspect. *Biochemistry (Moscow)* **82**, 1055–1060

(2017).
45. Metges, C. C. Contribution of Microbial Amino Acids to Amino Acid Homeostasis of the Host. *J. Nutr.* **130**, 1857S-1864S (2000).
46. Hartiala, J. A. *et al.* Genome-wide association study and targeted metabolomics identifies sex-specific association of CPS1 with coronary artery disease. *Nat. Commun.* **7**, 2705–2710 (2016).
47. Krall, A. S., Xu, S., Graeber, T. G., Braas, D. & Christofk, H. R. Asparagine promotes cancer cell proliferation through use as an amino acid exchange factor. *Nat. Commun.* **7**, 2705–2710 (2016).
48. Lange, L. A. *et al.* Genome-wide association study of homocysteine levels in Filipinos provides evidence for CPS1 in women and a stronger MTHFR effect in young adults. *Hum. Mol. Genet.* **19**, 2050–2058 (2010).
49. Kikuchi, G., Motokawa, Y., Yoshida, T. & Hiraga, K. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. *Proc. Jpn. Acad. Ser. B. Phys. Biol. Sci.* **84**, 246–63 (2008).
50. Gall, W. E. *et al.* alpha-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a nondiabetic population. *PLoS One* **5**, e10883 (2010).
51. Gantner, M. L. *et al.* Serine and Lipid Metabolism in Macular Disease and Peripheral Neuropathy. *N. Engl. J. Med.* NEJMoa1815111 (2019). doi:10.1056/NEJMoa1815111
52. Lone, M. A., Santos, T., Alecu, I., Silva, L. C. & Hornemann, T. 1-Deoxysphingolipids. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **1864**, 512–521 (2019).
53. Young, M. M., Kester, M. & Wang, H. G. Sphingolipids: Regulators of crosstalk between apoptosis and autophagy. *Journal of Lipid Research* **54**, 5–19 (2013).
54. Strettoi, E. *et al.* Inhibition of ceramide biosynthesis preserves photoreceptor structure and function in a mouse model of retinitis pigmentosa. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 18706–18711 (2010).
55. Clemons, T. E. *et al.* Medical characteristics of patients with macular telangiectasia type 2 (MacTel Type 2) MacTel Project Report No. 3. *Ophthalmic Epidemiol.* **20**, 109–113 (2013).
56. Meikle, P. J. & Summers, S. A. Sphingolipids and phospholipids in insulin resistance and related metabolic disorders. *Nat. Rev. Endocrinol.* **13**, 79–91 (2017).
57. Guasch-Ferré, M. *et al.* Metabolomics in prediabetes and diabetes: A systematic review and meta-analysis. *Diabetes Care* **39**, 833–846 (2016).
58. Meikle, P. J. *et al.* Plasma Lipid Profiling Shows Similar Associations with Prediabetes and Type 2 Diabetes. *PLoS One* **8**, 2705–2710 (2013).
59. Othman, A. *et al.* Plasma deoxysphingolipids: a novel class of biomarkers for the metabolic syndrome? *Diabetologia* **55**, 421–31 (2012).
60. Mwinyi, J. *et al.* Plasma 1-deoxysphingolipids are early predictors of incident type 2 diabetes mellitus. *PLoS One* **12**, 2705–2710 (2017).
61. Othman, A. *et al.* Lowering plasma 1-deoxysphingolipids improves neuropathy in diabetic rats. *Diabetes* **64**, 1035–1045 (2015).
62. Tan, S. Z., Begley, P., Mullard, G., Hollywood, K. A. & Bishop, P. N. Introduction to metabolomics and its applications in ophthalmology. *Eye* **30**, 773–783 (2016).
63. Laíns, I. *et al.* Metabolomics in the study of retinal health and disease. *Prog. Retin. Eye Res.* **69**, 57–79 (2019).
64. Laíns, I. *et al.* Human plasma metabolomics in age-related macular degeneration (AMD) using nuclear magnetic resonance spectroscopy. *PLoS One* **12**, 1–18 (2017).
65. Laíns, I. *et al.* Human Plasma Metabolomics Study across All Stages of Age-Related Macular Degeneration Identifies Potential Lipid Biomarkers. in *Ophthalmology*

125, 245–254 (2018).
66. Burgess, L. G. *et al.* Metabolome-wide association study of primary open angle glaucoma. *Investig. Ophthalmol. Vis. Sci.* **56**, 5020–5028 (2015).
67. Li, X., Luo, X., Lu, X., Duan, J. & Xu, G. Metabolomics study of diabetic retinopathy using gas chromatography-mass spectrometry: A comparison of stages and subtypes diagnosed by Western and Chinese medicine. *Mol. Biosyst.* **7**, 2228–2237 (2011).

Figure Legends

Fig. 1: Visual representation of all 121 significantly differentially abundant metabolites, comparing MacTel patients against controls. Each row represents a metabolite. The x-axis represents the LogFC. Negative LogFC values indicate reduced metabolite levels in MacTel patients compared to controls, and positive LogFC values indicate increased levels in MacTel patients. The model results are presented as dots indicating the estimated logFC with 95% confidence interval bars. Metabolites are divided into coloured blocks by their metabolic group. Mtb=Metabolism

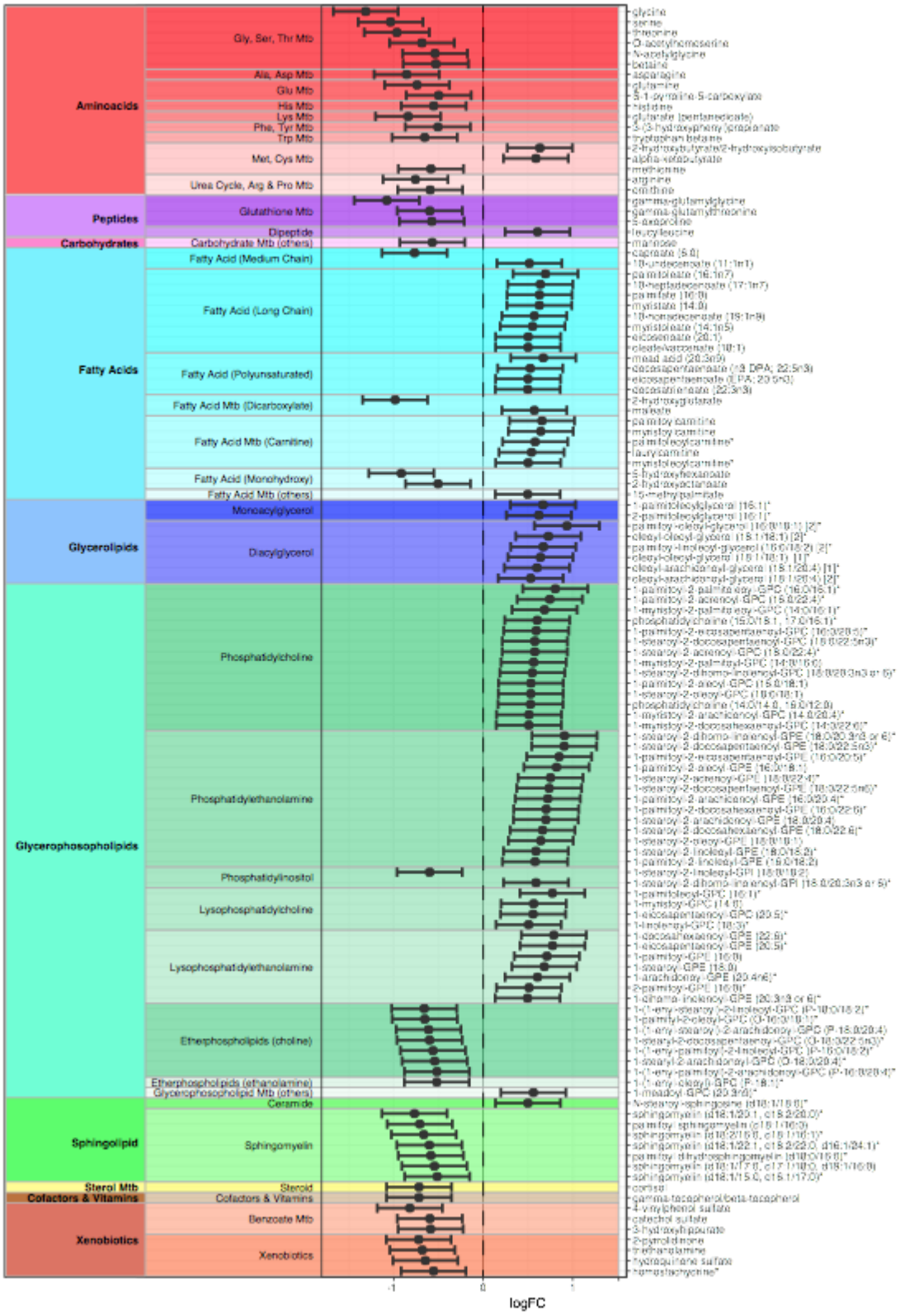
Fig. 2: Changes in abundance of 738 metabolites across the 50 metabolic groups. Each row represents a metabolic group. Significantly enriched metabolic groups are labelled with * for $p<0.05$ (corrected for FDR). Each group row is composed of blocks representing metabolites contained in the group. The colour of each block represents the Log_2 Fold-Change (logFC) of that metabolite comparing patients against controls. The colour blue represents depletion and magenta represents increased abundance. Mtb=Metabolism

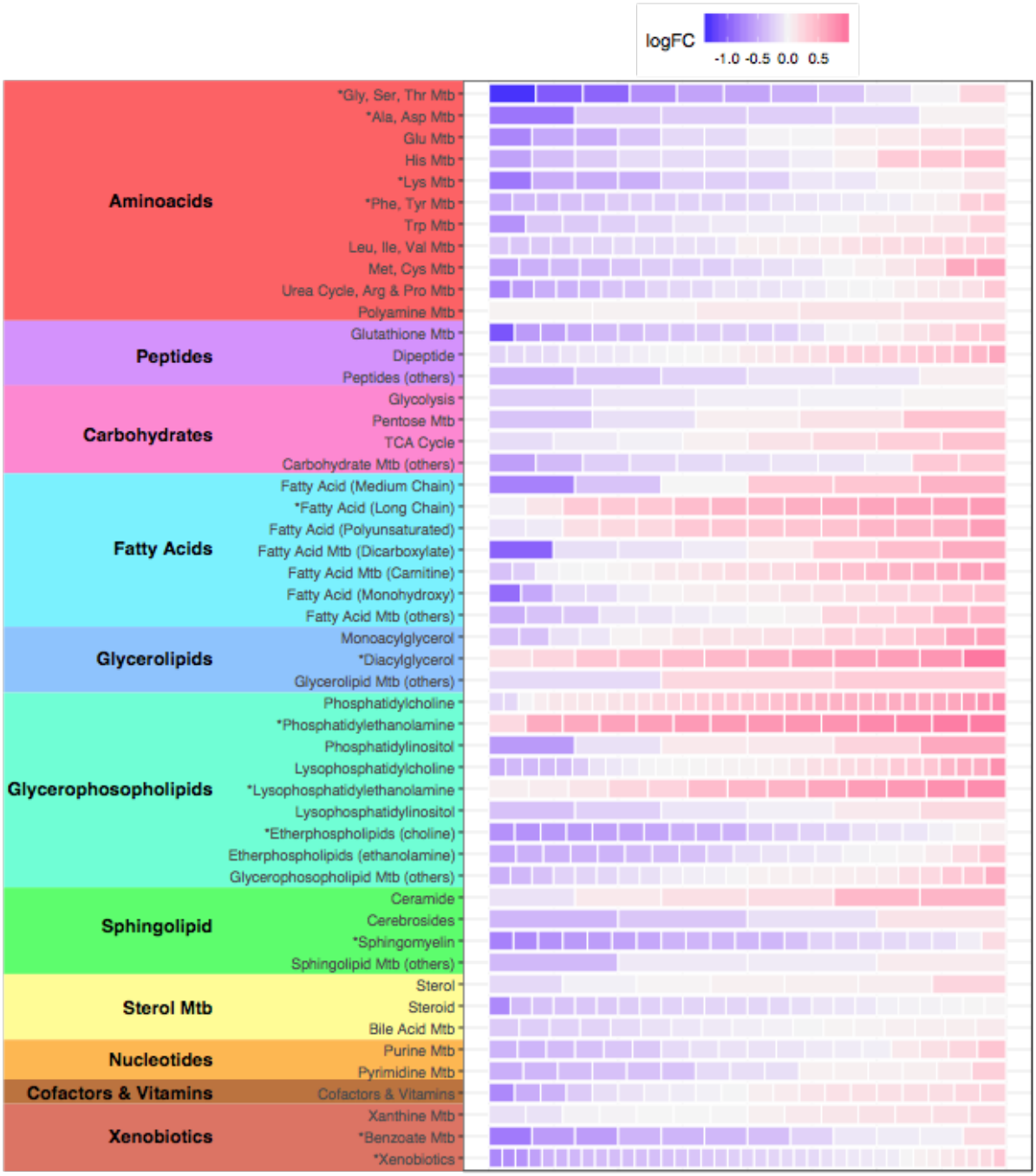
Fig. 3: Graphical overview of the key metabolic pathways that were affected in MacTel patients. Metabolites in blue were reduced in patients ($p<0.05$ in dark blue, $0.05<p<0.1$ in light blue). Metabolites in red were increased in MacTel patients ($p<0.05$ in dark red, $0.05<p<0.1$ in light red). Grey indicates no change between patients and controls, and metabolites on a white background were not measured. Double borders around metabolites indicates multiple metabolites within a group. The gene names of enzymes mentioned in the text are in yellow ovals. Note the generally reduced metabolite levels in glycine-serine and adjacent metabolic pathways, and generally increased levels in glycerophospholipid metabolism.

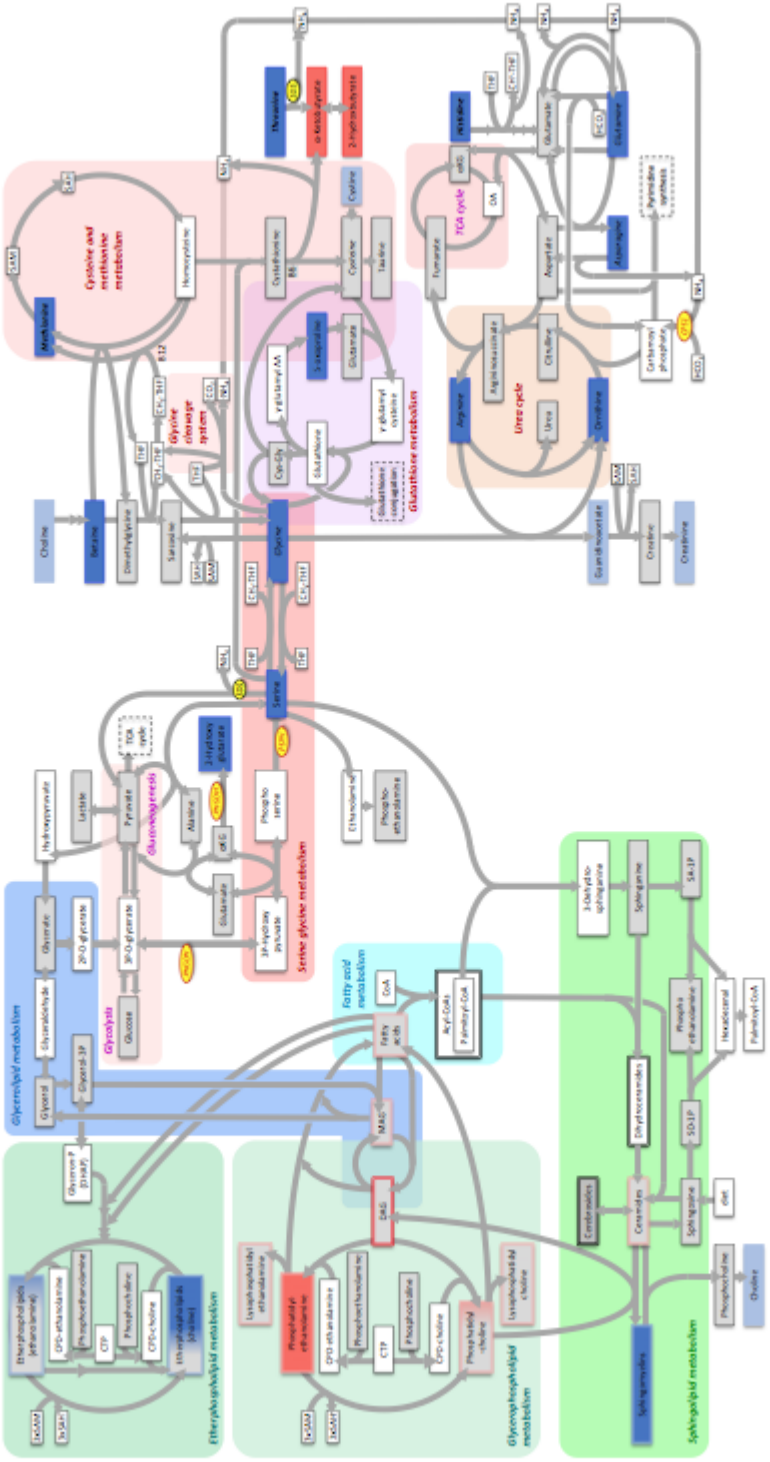
Fig. 4: Circos plot displaying 152 significantly differentially co-abundant metabolite pairs across 46 metabolic groups. Differential co-abundance between metabolic groups is represented by a line connecting the relevant groups. Thickness indicates the number of significant differential co-abundances. Correlations that were lost in patients are displayed in blue (positive correlation in controls and correlation lost in patients) and cyan (negative correlation in controls and correlation lost in patients). Newly formed connections in patients are presented in red (positive correlation in patients not observed in controls) and magenta

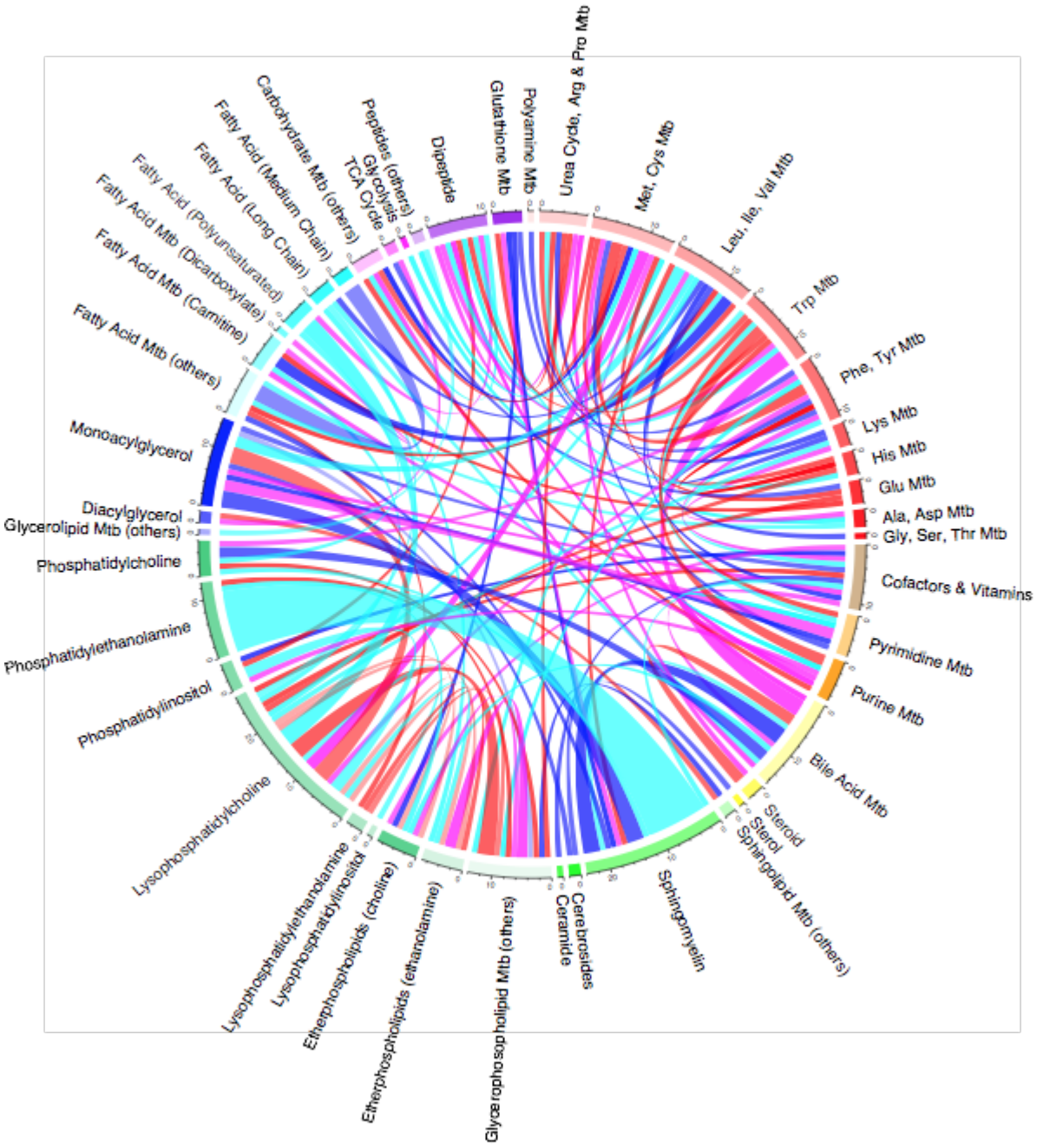
(negative correlation in patients not observed in controls). The transparency of the lines represents correlation magnitude (more transparency = lower magnitude). Note that connections involving the sphingomyelin group are most strongly suppressed in patients. Mtb=Metabolism

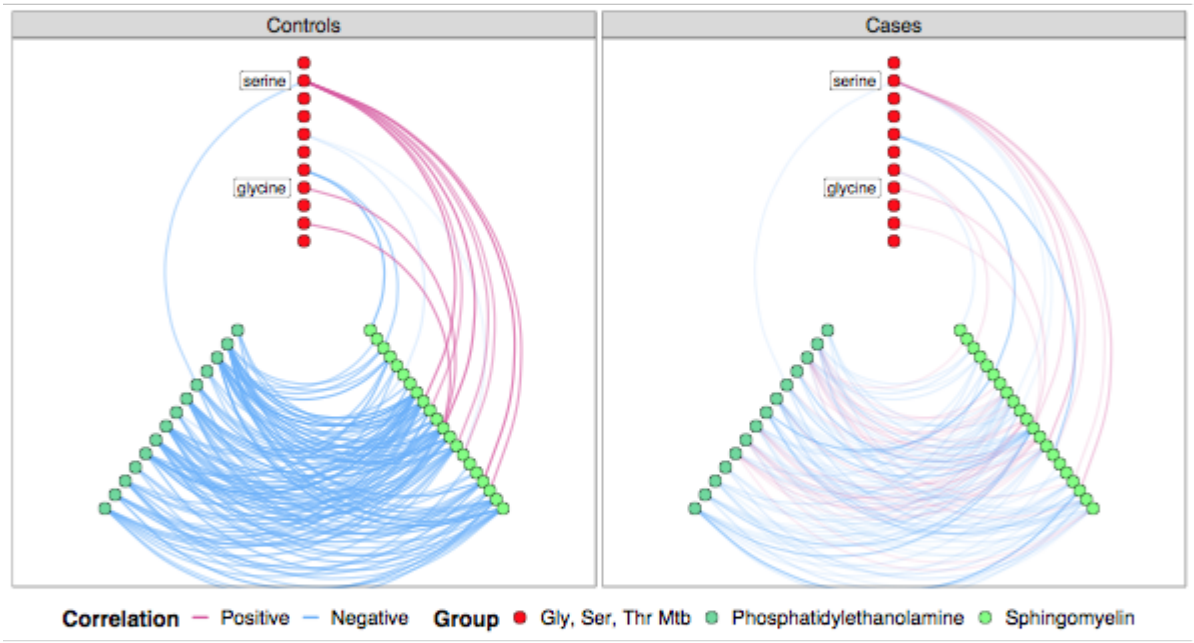
Fig. 5: Hive plots comparing co-abundance between metabolites in the sphingomyelin (13), phosphatidylethanolamine (14) and glycine-serine metabolism (11) groups in controls and MacTel patients. Specific metabolites are represented by circles. Co-abundance correlation between metabolites is represented by a line. Line transparency represents the correlation magnitude. Red lines represent positive correlations while blue lines represent negative correlations. Note that the majority of connections evident in controls are lost or markedly reduced in MacTel patients. Mtb=Metabolism


Aminoacids
Gly, Ser, Thr Mtb
Ala, Asp Mtb
Glu Mtb
His Mtb
Lys Mtb
Phe, Tyr Mtb
Trp Mtb
Met, Cys Mtb
Urea Cycle, Arg & Pro Mtb
Peptides
Glutathione Mtb
Dipeptide
Carbohydrates
Carbohydrate Mtb (others)
Fatty Acids
Fatty Acid (Medium Chain)
Fatty Acid (Long Chain)
Fatty Acid (Polyunsaturated)
Fatty Acid Mtb (Dicarboxylate)
Fatty Acid Mtb (Carnitine)
Fatty Acid (Monohydroxy)
Fatty Acid Mtb (others)
Glycerolipids
Monoacylglycerol
Diacylglycerol
Glycerophosospholipids
Phosphatidylcholine
Phosphatidylethanolamine
Phosphatidylinositol
Lysophosphatidylcholine
Lysophosphatidylethanolamine
Etherphospholipids (choline)
Etherphospholipids (ethanolamine)
Glycerophospholipid Mtb (others)
Sphingolipid
Ceramide
Sphingomyelin
Sterol Mtb
Sterol
Cofactors & Vitamins
Cofactors & Vitamins
Xenobiotics
Benzoate Mtb
Xenobiotics
glycine
serine
threonine
O-acetylhomoserine
N-acetylglycine
betaine
asparagine
glutamine
5-L-pyrroline-5-carboxylate
histidine
glutarate (pentanedioate)
3-(3-hydroxyphenyl)propanoate
tryptophan betaine
2-hydroxybutyrate/2-hydroxyisobutyrate
alpha-ketoglutarate
methionine
arginine
ornithine
gamma-glutamylglycine
gamma-glutamylthreonine
5-oxoproline
leucylleucine
mannose
caprate (10:0)
10-undecenoate (11:1n1)
palmitoleate (16:1n7)
10-heptadecenoate (17:1n7)
palmitate (16:0)
myristate (14:0)
10-nonadecenoate (19:1n9)
myristoleate (14:1n5)
eicosenoate (20:1)
oleate/vaccenate (18:1)
mead acid (20:3n9)
docosapentaenoate (n-3 DPA; 22:5n3)
eicosapentaenoate (EPA; 20:5n3)
docosatrienoate (22:3n3)
2-hydroxyglutarate
myristate
palmitoylcarnitine
myristoylcarnitine
palmitoleoylcarnitine*
laurylcarnitine
myristoleoylcarnitine*
5-hydroxyhexanoate
2-hydroxyoctanoate
15-methylpalmitate
1-palmitoleoylglycerol (16:1)*
2-palmitoleoylglycerol (16:1)*
palmitoyl-oleoyl-glycerol (16:0/18:1) [2]*
oleoyl-oleoyl-glycerol (18:1/18:1) [2]*
palmitoyl-linoleoyl-glycerol (16:0/18:2) [2]*
oleoyl-oleoyl-glycerol (18:1/18:1) [1]*
oleoyl-arachidonoyl-glycerol (18:1/20:4) [1]*
oleoyl-arachidonoyl-glycerol (18:1/20:4) [2]*
1-palmitoyl-2-palmitoleoyl-GPC (16:0/16:1)*
1-palmitoyl-2-adrenoyl-GPC (16:0/22:4)*
1-myristoyl-2-palmitoyl-GPC (14:0/16:1)*
phosphatidylcholine (16:0/16:1, 17:0/16:1)*
1-palmitoyl-2-eicosapentaenoyl-GPC (16:0/20:5)*
1-stearoyl-2-docosapentaenoyl-GPC (18:0/22:5n3)*
1-stearoyl-2-adrenoyl-GPC (18:0/22:4)*
1-myristoyl-2-palmitoyl-GPC (14:0/16:0)
1-stearoyl-2-dihomo-linolenoyl-GPC (18:0/20:3n3 or 6)*
1-palmitoyl-2-oleoyl-GPC (16:0/18:1)
1-stearoyl-2-oleoyl-GPC (18:0/18:1)
phosphatidylcholine (14:0/14:0, 16:0/12:0)
1-myristoyl-2-arachidonoyl-GPC (14:0/20:4)*
1-myristoyl-2-docosahexaenoyl-GPC (14:0/22:6)*
1-stearoyl-2-dihomo-linolenoyl-GPE (18:0/20:3n3 or 6)*
1-stearoyl-2-docosapentaenoyl-GPE (18:0/22:5n3)*
1-palmitoyl-2-eicosapentaenoyl-GPE (16:0/20:5)*
1-palmitoyl-2-oleoyl-GPE (16:0/18:1)
1-stearoyl-2-adrenoyl-GPE (18:0/22:4)*
1-stearoyl-2-docosapentaenoyl-GPE (18:0/22:5n6)*
1-palmitoyl-2-arachidonoyl-GPE (16:0/20:4)*
1-palmitoyl-2-docosahexaenoyl-GPE (16:0/22:6)*
1-stearoyl-2-arachidonoyl-GPE (18:0/20:4)
1-stearoyl-2-docosahexaenoyl-GPE (18:0/22:6)*
1-stearoyl-2-oleoyl-GPE (18:0/18:1)
1-stearoyl-2-linoleoyl-GPE (18:0/18:2)*
1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)
1-stearoyl-2-linoleoyl-GPI (18:0/18:2)
1-stearoyl-2-dihomo-linolenoyl-GPI (18:0/20:3n3 or 6)*
1-palmitoleoyl-GPC (16:1)*
1-myristoyl-GPC (14:0)
1-eicosapentaenoyl-GPC (20:5)*
1-linoleoyl-GPC (18:2)*
1-docosahexaenoyl-GPE (22:6)*
1-eicosapentaenoyl-GPE (20:5)*
1-palmitoyl-GPE (16:0)
1-stearoyl-GPE (18:0)
1-arachidonoyl-GPE (20:4n6)*
2-palmitoyl-GPE (16:0)*
1-dihomo-linolenoyl-GPE (20:3n3 or 6)*
1-(1-enyl-stearoyl)-2-linoleoyl-GPC (P-18:0/18:2)*
1-palmityl-2-oleoyl-GPC (O-16:0/18:1)*
1-(1-enyl-stearoyl)-2-arachidonoyl-GPC (P-18:0/20:4)
1-stearyl-2-docosapentaenoyl-GPC (O-18:0/22:5n3)*
1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2)*
1-stearyl-2-arachidonoyl-GPC (O-18:0/20:4)*
1-(1-enyl-palmitoyl)-2-arachidonoyl-GPC (P-16:0/20:4)*
1-(1-enyl-oleoyl)-GPC (P-18:1)*
1-myristoyl-GPC (20:3n6)*
N-stearoyl-sphingosine (d18:1/18:0)*
sphingomyelin (d18:1/20:1, c18:2/20:0)*
palmitoyl sphingomyelin (d18:1/16:0)
sphingomyelin (d18:2/16:0, c18:1/16:1)*
sphingomyelin (d18:1/22:1, c18:2/22:0, d16:1/24:1)*
palmitoyl dihydrosphingomyelin (d18:0/16:0)*
sphingomyelin (d18:1/17:0, c17:1/18:0, d18:1/16:0)
sphingomyelin (d18:1/15:0, c15:1/17:0)*
cortisol
gamma-tocopherol/beta-tocopherol
4-vinylphenol sulfate
catechol sulfate
3-hydroxyhippurate
2-pyrrolidinone
1-methylnicotinamide
hydroquinone sulfate
*
-1
0
1
logFC


logFC
-1.0 -0.5 0.0 0.5
Aminoacids
*Gly, Ser, Thr Mtb
*Ala, Asp Mtb
Glu Mtb
His Mtb
*Lys Mtb
*Phe, Tyr Mtb
Trp Mtb
Leu, Ile, Val Mtb
Met, Cys Mtb
Urea Cycle, Arg & Pro Mtb
Polyamine Mtb
Peptides
Glutathione Mtb
Dipeptide
Peptides (others)
Carbohydrates
Glycolysis
Pentose Mtb
TCA Cycle
Carbohydrate Mtb (others)
Fatty Acids
Fatty Acid (Medium Chain)
*Fatty Acid (Long Chain)
Fatty Acid (Polyunsaturated)
Fatty Acid Mtb (Dicarboxylate)
Fatty Acid Mtb (Carnitine)
Fatty Acid (Monohydroxy)
Fatty Acid Mtb (others)
Glycerolipids
Monoacylglycerol
*Diacylglycerol
Glycerolipid Mtb (others)
Glycerophosopholipids
Phosphatidylcholine
*Phosphatidylethanolamine
Phosphatidylinositol
Lysophosphatidylcholine
*Lysophosphatidylethanolamine
Lysophosphatidylinositol
*Etherphospholipids (choline)
Etherphospholipids (ethanolamine)
Glycerophosopholipid Mtb (others)
Sphingolipid
Ceramide
Cerebrosides
*Sphingomyelin
Sphingolipid Mtb (others)
Sterol Mtb
Sterol
Steroid
Bile Acid Mtb
Nucleotides
Purine Mtb
Pyrimidine Mtb
Cofactors & Vitamins
Cofactors & Vitamins
Xenobiotics
Xanthine Mtb
*Benzoate Mtb
*Xenobiotics


Cysteine and methionine metabolism
Glycine cleavage system
Glutathione metabolism
TCA cycle
Urea cycle
Serine glycine metabolism
Glycolysis
Gluconeogenesis
Fatty acid metabolism
Glycerolipid metabolism
Etherphospholipid metabolism
Glycerophospholipid metabolism
Sphingolipid metabolism


Glycerophospholipid Mtb (others)
Ether phospholipids (ethanolamine)
Ether phospholipids (choline)
Lysophosphatidylinositol
Lysophosphatidylethanolamine
Lysophosphatidylcholine
Phosphatidylinositol
Phosphatidylethanolamine
Phosphatidylcholine
Glycerolipid Mtb (others)
Diacylglycerol
Monoacylglycerol
Fatty Acid Mtb (others)
Fatty Acid Mtb (Carnitine)
Fatty Acid Mtb (Dicarboxylate)
Fatty Acid (Polyunsaturated)
Fatty Acid (Long Chain)
Fatty Acid (Medium Chain)
Carbohydrate Mtb (others)
Peptides (others)
Glycolysis
TCA Cycle
Dipeptide
Glutathione Mtb
Polyamine Mtb
Urea Cycle, Arg & Pro Mtb
Met, Cys Mtb
Leu, Ile, Val Mtb
Trp Mtb
Phe, Tyr Mtb
Lys Mtb
His Mtb
Glu Mtb
Ala, Asp Mtb
Gly, Ser, Thr Mtb
Cofactors & Vitamins
Pyrimidine Mtb
Purine Mtb
Bile Acid Mtb
Sterol
Sterol
Sphingolipid Mtb (others)
Sphingomyelin
Ceramide
Cerebrosides


Controls
Cases
serine
glycine
serine
glycine
Correlation
Positive
Negative
Group
Gly, Ser, Thr Mtb
Phosphatidylethanolamine
Sphingomyelin

4.3 Discussion

The work presented above confirmed the impact on the metabolome of MacTel patients due to the genetically induced depletion of glycine and serine that these patients harbour. The updated schematic of the study findings is presented in **Figure 31**. However, the analysis also revealed novel risk factors for the disease. Among these, we found that strong dysregulation of sphingolipids, as well as phosphatidylethanolamines over-abundance, are likely to be part of the disease risk. The study also found interesting overlaps between the metabolic biomarkers of MacTel disease with those of type 2 diabetes.

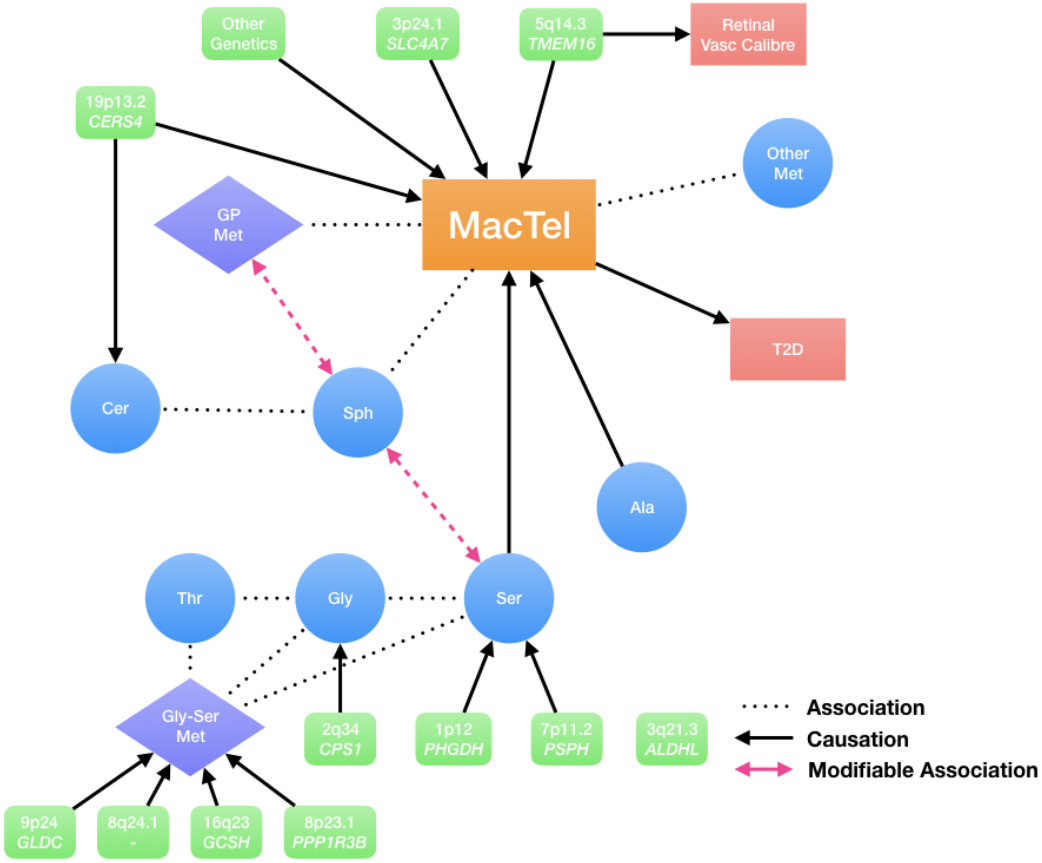


Figure 31: Main findings schematics displaying the drivers and traits associated with MacTel. Genetic traits are displayed as green rounded squares. Metabolites are displayed as blue circles. Metabolic pathways are displayed as purple diamonds. Phenotypic traits are displayed as red squares. Associations are represented by dotted black lines while associations changed by interaction are displayed by dashed red arrows. Causality is indicated by unidirectional solid black arrows.

5 Contribution to Gantner et al. Serine and Lipid Metabolism Link Macular Disease and Peripheral Neuropathy

5.1 Introduction

This chapter describes an investigation into the role of serine depletion and subsequent deoxy-sphingolipids accumulation on MacTel disease which will be published by the New England Journal of Medicine this year (83). RB performed several statistical investigations for this research. This chapter will first introduce the rationale behind the study and introduce the main findings, which were obtained without RB's contribution, followed by a rationale for the analyses directly performed by RB. Then, methods, results and discussion of analyses performed by RB and reported in Gantner et al, and related unpublished analyses will be presented.

5.1.1 Serine depletion might induce deoxy-sphingolipid accumulations

In the previous chapters, we described how genetic influences affecting MacTel patients are likely to induce metabolic disturbances which in turn are likely to be a key contributor to MacTel aetiology. Amongst these observations, we found that MacTel is likely caused by serine depletion as well as alanine overabundance. As mentioned in Chapter 3, serine depletion might be compensated for by glycine. However, we have observed MacTel patients having genetic variants which prevent glycine from being biosynthesised correctly, hence impeding its conversion into serine.

The biological mechanism through which serine depletion may cause MacTel is largely unknown, and is likely complex, given the central role that serine plays in several metabolic pathways. Serine is an essential amino acid for a specific pathway which is central for this chapter called the sphingolipid pathway. In fact, serine is central for the biosynthesis of all sphingolipids requiring a sphinganine base. To synthesise sphinganine (SA), the enzyme palmitoyltransferase (SPT) needs to condense serine with palmitoyl-CoA (84). However, when serine is depleted, the SPT enzyme may condense palmitoyl-CoA with alanine (85). The SA created by this process lacks the hydroxyl group that differentiates serine from alanine. This alternative compound, deoxy-sphinganine (doxSA) generates subsequent deoxy-sphingolipids (doxSL). A schematic from Gantner et al (83) illustrating SPT promiscuity is presented in **Figure 32**.

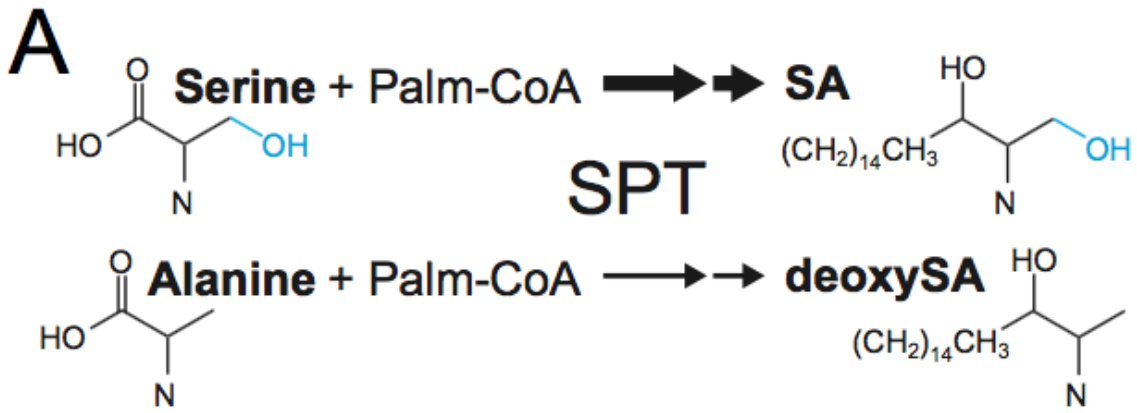


Figure 32: Schematic from Gantner et al depicting different sphinganines resulting from the reaction of palmitoyl-CoA with serine or alanine (83).

DoxSL have been previously associated with different medical conditions such as several metabolic syndromes, type 2 diabetes, liver syndromes and other serine deficiency disorders (83). DoxSL lipids have also been observed to have toxic effects on different cell types, specifically neurons, including retinal photoreceptors.

5.1.2 Hereditary Sensory Neuropathy Type 1

SPT is a trimer formed from the products of three genes; *SPTLC1*, *SPTLC2*, and *SPTLC3* (86). Rare coding mutations in the first two of these genes have been reported in patients suffering from a peripheral neuropathy disease called Hereditary sensory neuropathy type 1 (HSAN1) (84). Specifically, these mutations have been described to modify the structure of the SPT enzyme, increasing its affinity for alanine. This aberration in affinity results in an accumulation of doxSL, even when the abundance of metabolic serine is within the normal

physiological range. An extract from Gantner et al showing this mechanism is presented in **Figure 33**.

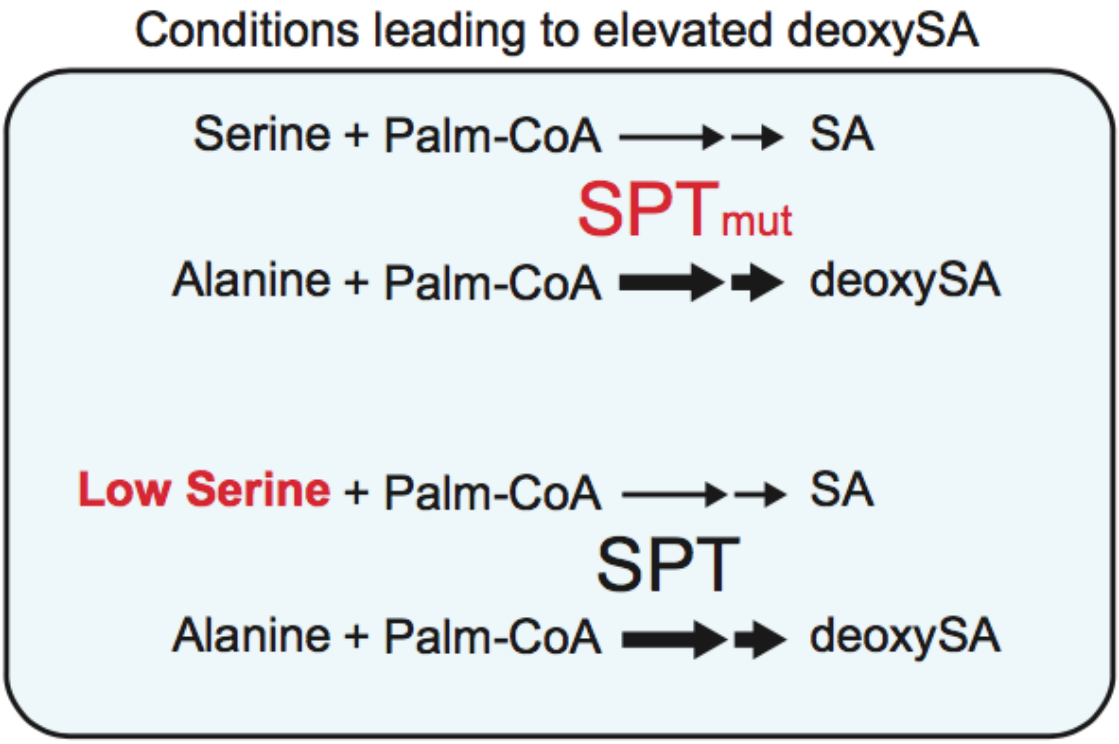


Figure 33: Schematic of conditions leading to elevated doxSA levels. The top scenario displays SPT condensing palmitoyl-CoA with alanine because of rare mutations affecting enzyme specificity. The bottom scenario displays SPT condensing palmitoyl-CoA with alanine because of lack of available serine. Both scenarios result in deoxy-sphinganine production (83).

A very early onset MacTel patient was observed, and whole exome sequencing revealed a mutation in *SPTLC1*. The same subject was also discovered to be affected by HSAN1 disease. This extraordinary finding together with the

functional similarity between the *SPTLC1* mutation and serine depletion provided the first insight into the metabolic basis of MacTel from a Mendelian disease perspective. Researchers from the MacTel consortium set out to investigate MacTel comorbidity among patients suffering from HSAN1.

5.1.3 HSAN1 patients carrying specific mutations are affected by MacTel disease

The authors of Gantner et al investigated mutations shared among first-degree relatives of the early-onset patient and found the father and a sibling of the original proband to be affected or "possibly affected" by MacTel disease, while the mother and another sibling were unaffected. All MacTel affected individuals were also affected by HSAN1 (initially misdiagnosed as Charcot-Marie-Tooth disease, another more common and similar neuropathy, with many known genes). WES analysis, performed by our MacTel collaborator Dr Rando Allikmets, Columbia University, New York, revealed all affected subjects to carry the p.Cys133Tyr mutation in the *SPTLC1* gene, as shown in **Figure 34**.

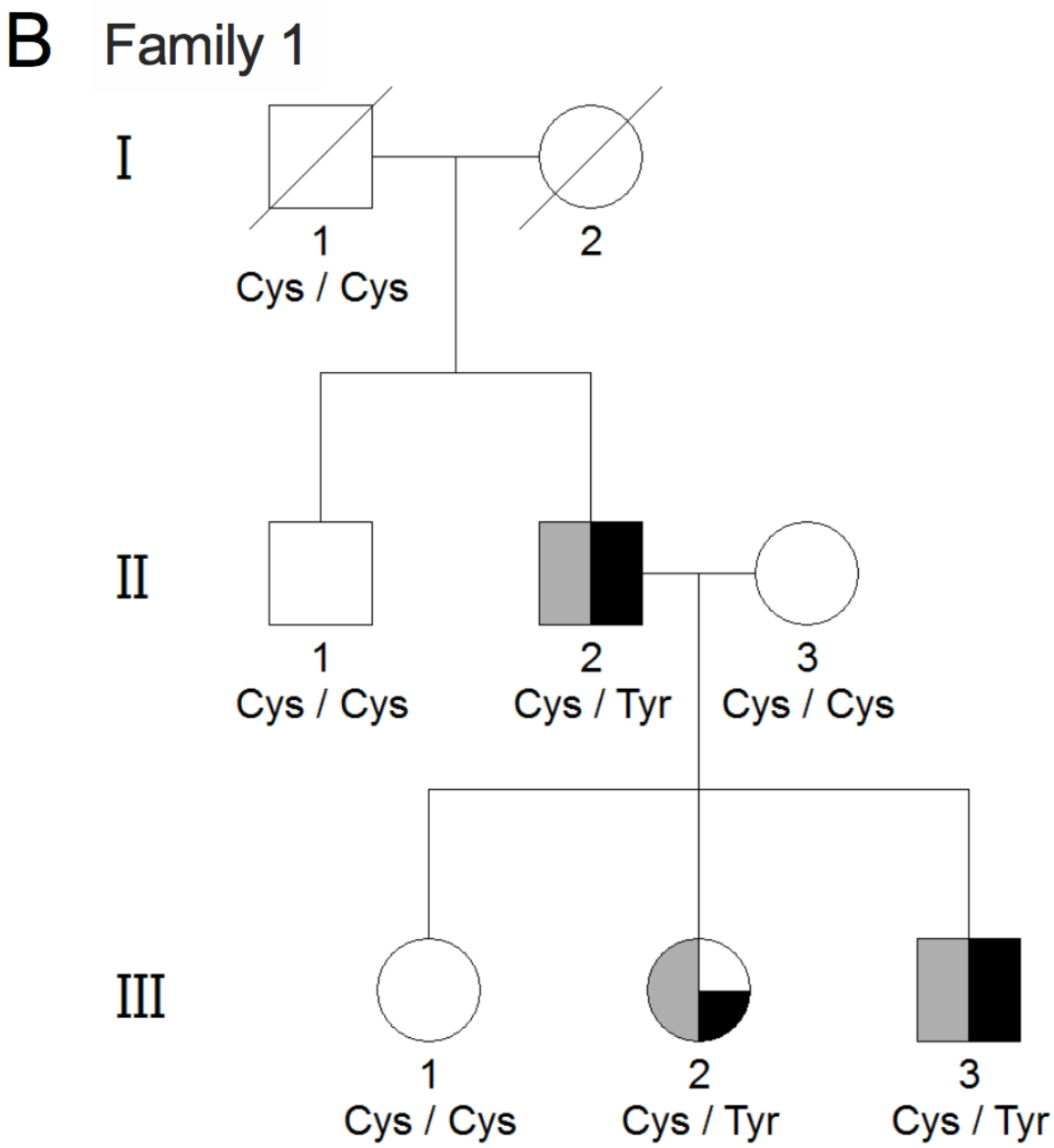


Figure 34: Extract from Gantner et al showing affected HASN1 individuals (grey) and affected MacTel (black) with their respective mutation in the SPTLC1 gene (83).

This mutation is specifically associated with HSAN1 subtype A. Due to this finding, 11 other individuals, also affected by HSAN1 from different families with no known relationship with the aforementioned patients were examined for

MacTel. Among these, 8 were affected, bringing the total to 11 out of 14 HSAN1 patients comorbid with MacTel. The enrichment of MacTel prevalence among this HSAN1 cohort was extremely significant, due to the known low MacTel prevalence ($p<2.2e-16$). Strikingly, of the three patients affected by HSAN1 disease but not MacTel, two presented with the known pathogenic p.Cys133Trp mutation in *SPTLC1,* while the remaining patient also had the p.Cys133Tyr mutation, but was young compared with the average age of MacTel onset, suggesting incomplete penetrance. Additionally, two of the patients affected by both MacTel and HSAN1 presented with the p.Ser384Phe mutation in *SPTLC2*.

5.1.4 Induced plasma serine depletion increases doxSL abundance in the mouse retina and causes nervous dysfunction

The relationship between circulating serine and levels of doxSL in the retina is unknown. To investigate whether the depletion of serum serine affects retinal levels of doxSL, our colleagues investigated retinal serine concentrations in mice fed a low glycine and serine diet. These mice had lower levels of circulating serine which was also associated with excessive circulating doxSA; the same was mirrored in the retina. Functional testing revealed that mice on the low glycine and serine diet developed retinal and sensory deficit after 10 months, mirroring the effect observed in MacTel patients (83).

5.1.5 Study aims: Exploring doxSA levels on MacTel patients as well as their relationship with other metabolites and disease progression

The aforementioned findings showed that lower serine, as well as rare mutations in genes encoding the SPT enzyme, produce greater doxSL levels, which in turn are likely to contribute to neurodegeneration in the retina, characteristic of MacTel. However, no direct study of the levels of doxSL in MacTel patients has ever been performed. Measuring doxSL levels on individuals suffering from MacTel as well as in healthy controls allows exploration of several hypotheses.

A. If serine depletion is associated with a higher abundance of doxSA, then MacTel patients should present with elevated doxSA compared to controls.
B. Among both MacTel patients and controls, the levels of circulating doxSA should correlate negatively with serine, and positively with alanine concentrations.
C. As we might expect doxSA accumulation to disrupt the function and turn-over of downstream lipid species, there might be differences in SA levels between MacTel patients and controls.
D. T2D is known comorbidity with MacTel and doxSA levels have been previously associated with type 2 diabetes. We may then expect that patients suffering from both MacTel and T2D might present different levels of doxSA from patients suffering from any of the two diseases alone.
E. If doxSA is important for the onset of MacTel, levels of this metabolite might also contribute to disease progression.

F. Identifying additional associations with other metabolites with doxSA levels untargeted metabolomics fashion might shed light on the more complex metabolic disturbances observed in previous chapters.

All the aforementioned hypotheses, excluding the last, were explored and published in Gantner et al, and all are further described in the following sections.

5.2 Methods

5.2.1 Analysis of deoxy-sphingolipids in MacTel

To test the aforementioned hypotheses we conducted a new, **targeted** metabolomics analysis. The targeted metabolites in this new study were, doxSA, SA, serine, alanine and 11 additional amino acids quantified in 125 unrelated MacTel patients and 95 unrelated controls by the Christian Metallo group at the University of California, San Diego (as detailed in Gantner et al). DoxSA and SA were measured in separate batches from the other metabolites. All subjects were confirmed not to have HSAN1 through WES sequencing. For all subjects, we collect information including sex at birth, ethnicity, age and T2D status. Geographical origin of the samples, as well as metabolomics analysis batch, were recorded to account for potential systematic variation in the data. Patients were assessed for Ellipsoid Zone loss (EZ loss) area which, as mentioned in Chapter 1, is a result of the loss of the mitochondria-rich photoreceptor inner segments and is a symptom of MacTel progression. To correct the skewed distribution of the

measured metabolomic abundances we applied a log2 transformation. Similarly, we applied a square root transformation to the skewed distribution of EZ loss area, obtaining a proxy for EZ loss radius from the centre.

To test for association between metabolomics levels with MacTel disease, EZ loss and T2D, as well the association between metabolites we adopted a linear mixed model regression strategy of the form

$$M_i = \beta x_i + \alpha C_i + \gamma_{iM} + e_i$$

Where $M_j$is the log2 abundance of metabolite M for a subject i measured in batch j, β is the fixed effect of the factor of interest x_i, α the vector of fixed effects of covariates measured in the vector C_i, and $\gamma_{iM}$is the random effect capturing the influence of batch used to measure on the abundance of metabolite M in subject i, e_i is the error term assumed to be normally distributed and centred around zero.

This method enabled us to test for association between different factors while at the same time taking into account all collected covariates as well as the batches used to perform metabolomics measurements. Given the imbalance between cases and controls among batches, we included batch information as a random effect rather than a fixed effect.

The hypothesis that DoxSA increment is a result of the co-presence of MacTel and T2D was tested by including an interaction term between MacTel status and T2D status using the model

$$doxSA_i = \beta_0 + \beta_1 MT_i + \beta_2 T2D_i + \beta_3 MT_i * T2D_i + \alpha C_i + \gamma_{doxsa(i)} + e_i$$

Where $doxSA_i$ is the doxSA abundance for subject i measured in batch $\gamma_{doxsa_{(i)}}$, $\beta_1$is the fixed effect of MacTel status for subject $i$measured through MT_i, $\beta_2$is the fixed effect of T2D status for subject $i$measured through $T2D_i$, $\beta_3$is the fixed effect of the interaction between MacTel and T2D status, $\beta x_i$is the contribution of additional covariates and $\gamma_{doxsa_{(i)}}$ is the random effect capturing the influence of the batch used to measure abundances of $doxSA$ in subject.

When testing for association between different metabolites, the batch of the metabolite was included as a dependent variable as a random effect while the estimated batch effect for the metabolite was included indirectly by adjusting the other metabolite for batch effect and including this residual instead of the original metabolite in a mixed model:

$$M1_i = \alpha C_i + \delta M2^r{}_i + \gamma_{m1i} + e_i$$

where $M1_i$ is the abundance of metabolite M1 measured using batch $m1$,and $\delta$is the effect of the residualised abundance $M2^r{}_i$ of metabolite M2 obtained by subtracting its own batch effect estimated prior to the model using the formula

$$M2^r{}_i = M2_i - \widehat{\gamma_{m2\,i}}$$

where $M2_i$ is the abundance of metabolite 2 and $\widehat{\gamma_{m2\,i}}$ is the best unbiased linear estimation of the random effect the effect of batch $m2$ used to measure metabolite $M2_i$ estimated through the following mixed model

$$M2_i = \alpha C_i + \gamma_{m2\,i} + e_i$$

Average and maximum EZ loss was calculated for all individuals, averaged across both eyes where binocular data was available. P-values were calculated using the Satterthwaite degrees of freedom as implemented by the R package LmerTest (87, 88).

5.2.2 Exploring untargeted metabolomic associations with deoxy-sphingolipids

Untargeted metabolomics association with doxSA levels were tested by combining the targeted metabolomics datasets used in this study with the cleaned untargeted metabolomics dataset presented in Chapter 4. Unfortunately, not all subjects from the targeted metabolomics study (125 cases and 95 controls) were also included in the initial untargeted metabolomics cohort (60 cases vs 58 controls). The combined dataset contained 739 metabolites measured for 93 subjects (55 cases and 38 healthy controls). The estimated batch effect on doxSA levels was extracted *a priori* from the original doxSA levels as described above and the association was tested using linear regression models, correcting for multiple covariates including

disease status, which may have a confounding effect on the association between metabolites. To this end, we used the model

$$doxSA^{r}{}_{i} = \alpha C_i + \delta \ M_i{}^{c} + e_i$$

where $doxSA^{r}{}_{i}$ is the batch-residualised abundance of $doxSA$, δ is the effect of the abundance of metabolite $M_i{}^{c}$measured in the data presented in Chapter 2 and Chapter 4 and cleaned with the same process presented in Chapter 4.

To test for independence of effects between different metabolites on doxSA levels we adopted a stepwise conditional modelling approach, whereby the most significantly associated metabolite was included in the model and the remainder were tested for associations. This process was iterated until no significant metabolites remained. Differences in the association between metabolites, given disease status, were tested using an interaction term between the metabolite of interest and the disease status using the following model

$$doxSA^{r}{}_{i} = \alpha C_i + \delta_1 M_i{}^{c} + \delta_2 MT_i + \delta_3 MT_i * M_i + e_i$$

where δ_2 is the effect of MacTel status MT_i and δ_3 is the interaction between MacTel status MT_i and abundance of metabolite M_i.

To explore whether metabolites were interacting with doxSA and modifying MacTel risk, we performed a logistic regression model including MacTel status as

the dependent variable and testing for interaction of each metabolite with doxSA using the following model

$$logit(P_{MacTel})_i = \alpha C_i + \delta_1 {M_i}^c + \delta_2 {doxSA^r}_i + \delta_3 {doxSA^r}_i * {M_i}^c + e_i$$

Where the $logit(P_{MacTel})_i$ is the probability of observing MacTel on the logit scale and $\delta_3$represents the interaction between metabolite $M$and doxSA levels.

To test whether some metabolites were responsible for the interaction effect between MacTel and TD2 on doxSA levels we used a triple interaction linear model defined as follows:

$$\begin{gathered} {doxSA^r}_i = \beta_0 + \beta_1 MT_i + \beta_2 T2D_i + \\ + \beta_3 MT_i * T2D_i + \beta_4 MT_i * {M_i}^c + \beta_5 T2D_i * {M_i}^c + \\ + \beta_6 MT_i * T2D_i * {M_i}^c + \alpha C_i + \gamma_{doxsa} + e_i \end{gathered}$$

Where β_6 measured the effect of metabolite M_i on the interaction between MacTel status MT_i and diabetes status $T2D_i$.

Entire pathway abundances were included in the regression models by performing Principal Component Analysis on a dataset containing only the metabolites of a particular metabolic pathway and extracting the first principal component as described in Chapter 4. To correct for false discovery arising from multiple testing we applied the Benjamini-Hochberg correction obtaining corrected p-values (P(fdr)).

5.3 Results

5.3.1 Analysis of deoxy-sphingolipids in MacTel

MacTel patients presented on average with an increment of 0.61 on the log scale of doxSA levels compared to the controls (p-value<1e-7). None of the covariates (T2D status, age, etc) were associated with doxSA levels when correcting for MacTel status.

As expected serine was significantly decreased in MacTel patients (β=-0.23, p<1e-5) and doxSA abundance negatively correlated with serine after correcting for disease status (δ=-1.28, p<1e-7). Interestingly, in the same model, MacTel effect on doxSA was also significant (β=0.32, p= 7.8e-5) indicating an over-abundance of doxSA in MacTel patients not explained by serine depletion. In a separate model, we found a significant positive association between alanine and doxSA (δ=0.64, p=4.7e-5), as well as a significant overabundance of doxSA in MacTel cases (β=0.53, p<1e-7). When serine, alanine and MacTel were included in the same regression model, we intriguingly found both amino acids to increase in significance (δ_{serine}=-1.36, p_{serine} <2e-16 ; $\delta_{alanine}$=0.78, $p_{alanine}$ <1e-9), as well as a concomitant decrease in MacTel status significance (β=0.18, p=0.013), suggesting that the relationship of these metabolites with doxSA was able to explain most of the relationship observed between MacTel status and doxSA.

When testing for differences in SA levels between MacTel cases and healthy controls while correcting for all possible covariates we found no significant effect of MacTel status on SA abundance (β =0.04, p=0.476). Similar results were obtained when removing all non-significant covariates from the model.

When testing the effect of T2D on doxSA concentrations in MT cases, we found higher doxSA in cases with T2D (β_3= 0.33, p-value=0.054). From the same model, we also observed MacTel patients without T2D to present higher doxSA levels when compared to controls without T2D (β_1= 0.52, p=5e-7). Interestingly, we did not find any significant difference between patients suffering from T2D alone compared with healthy controls (β_2=-11, p=0.407 - **Figure 39**).

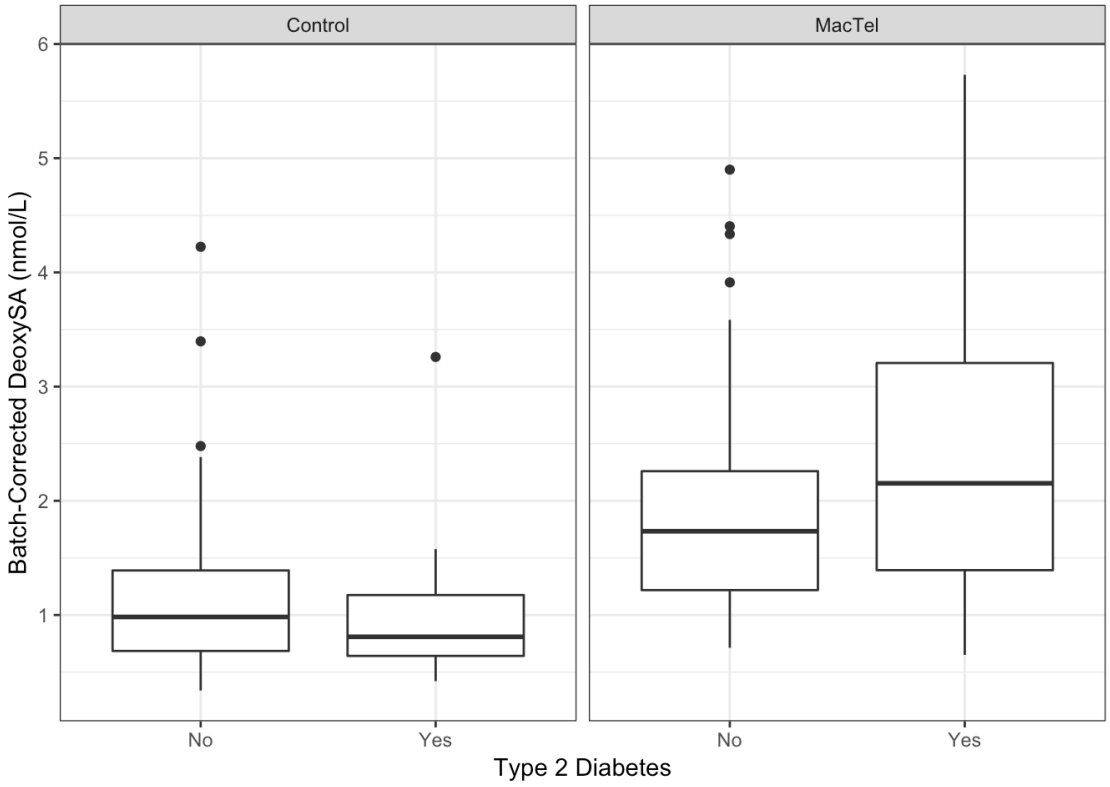


Figure 35: Distribution differences in batch-corrected doxSA levels between MacTel cases and controls shown according to type 2 diabetes status.

The relationship between EZ loss radius and doxSA levels was tested using 117 MacTel patients for whom EZ loss data was available. Maximum EZ loss (i.e. from the most affected eye) and mean loss across both eyes were positively associated with doxSA levels (β=0.14, p=0.046; **Figure 36**; and β=0.155, p=0.064 respectively).

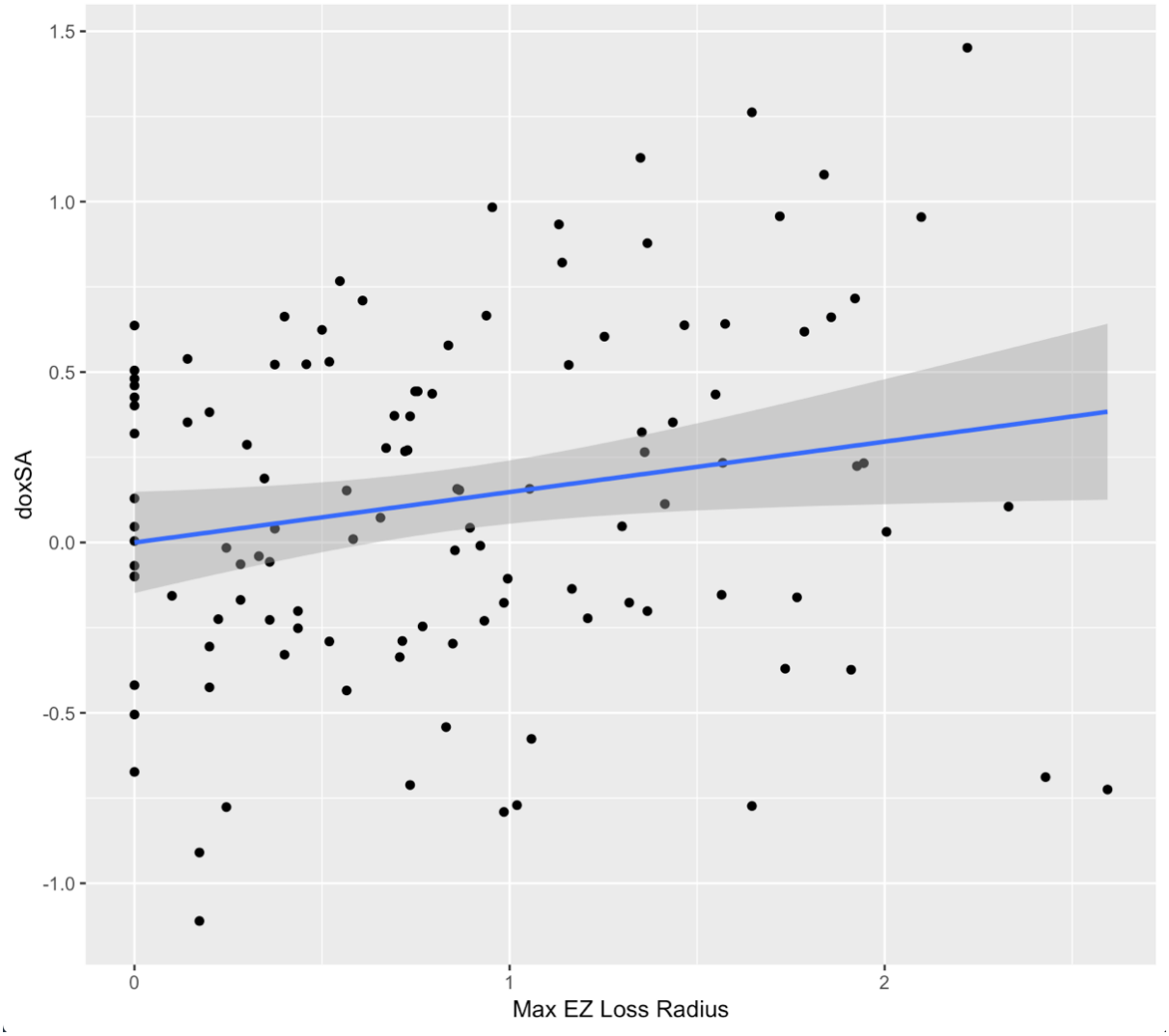


Figure 36: Relationship between maximum EZ loss lesion and corrected log2 doxSA levels.

5.3.2 Exploring untargeted metabolomic associations with deoxy-sphingolipid abundances

We tested for association between all metabolite abundances and doxSA levels. To distinguish between metabolites, present in both datasets, we will refer to them as either "new" (measured in targeted metabolomics study) or "original" (measured in untargeted metabolomics study). We found age and MacTel status to be necessary covariates in the model when using the combined targeted and untargeted metabolomics dataset.

We identified the original serine concentration (serine_original) to be the most significantly associated metabolite with doxSA (δ=-0.49, P(fdr)=9.4e-5). The new serine concentrations (serine_new), showed a more modest association with doxSA (δ-0.34, P(fdr)=0.0157). New SA levels (sphinganine_new), were the second-most

associated with doxSA (δ=0.38, P(fdr)=0.0064). All significantly associated metabolites are provided in **Table 1**.

Metabolite	Pathway	Effect size	p-value	P(fdr)
serine_original	Gly, Ser, Thr Mtb	-0.49	1.0E-07	9.4E-05
sphinganine_new	Sphingolipids Mtb	0.38	1.7E-05	0.006
1-stearoyl-2-adrenoyl-GPE (18:0/22:4)*	Phosphatidylethanolamine	0.38	4.0E-05	0.008
1-palmityl-2-oleoyl-GPC (O-16:0/18:1)*	Plasmenylcholin	-0.37	4.1E-05	0.008
1-stearoyl-2-docosapentaenoyl-GPE (18:0/22:5n6)*	Phosphatidylethanolamine	0.37	7.1E-05	0.011
serine_new	Gly, Ser, Thr Mtb	-0.34	0.0001	0.016
1-(1-enyl-palmitoyl)-2-docosahexaenoyl-GPC (P-16:0/22:6)*	Plasmenylcholin	-0.31	0.0003	0.031
gamma-glutamylisoleucine*	Glutathione Mtb	0.31	0.0003	0.031
threonine	Gly, Ser, Thr Mtb	-0.34	0.0005	0.032
lactosyl-N-nervonoyl-sphingosine (d18:1/24:1)*	Cerebrosides	-0.30	0.0006	0.032
taurolithocholate 3-sulfate	Bile Acid Mtb	-0.29	0.0006	0.032

erythritol	Xenobiotics	0.30	0.0006	0.032
allantoin	Purine Mtb	0.30	0.0006	0.032
1-nervonoyl-GPC (24:1n9)*	Lysophosphatidylcholine	-0.29	0.0006	0.032
arabitol/xylitol	Pentose Mtb	0.30	0.0006	0.032
1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)*	Plasmenylcholin	-0.30	0.0007	0.032
1-adrenoyl-GPC (22:4)*	Lysophosphatidylcholine	0.30	0.0007	0.032
docosahexaenoate (DHA; 22:6n3)	Fatty Acid (Polyunsaturated)	-0.29	0.0008	0.032
2-hydroxy-3-methylvalerate	Leu, Ile, Val Mtb	0.29	0.0009	0.032
N-acetyltaurine	Met, Cys Mtb	0.29	0.0009	0.032
1-(1-enyl-stearoyl)-2-docosahexaenoyl-GPC (P-18:0/22:6)*	Plasmenylcholin	-0.29	0.0009	0.032
N-palmitoyl-sphinganine (d18:0/16:0)	Ceramide	0.32	0.0009	0.032
1-methylxanthine	Xanthine Mtb	-0.28	0.0012	0.040
1-eicosenoyl-GPC (20:1)*	Lysophosphatidylcholine	-0.28	0.0013	0.040
5-methyluridine (ribothymidine)	Pyrimidine Mtb	-0.28	0.0013	0.041
valylarginine	Dipeptide	-0.27	0.0015	0.045
5-(galactosylhydroxy)-L-lysine	Lys Mtb	0.28	0.0017	0.047
alpha-hydroxyisovalerate	Leu, Ile, Val Mtb	0.28	0.0018	0.047
N-acetylproline	Urea Cycle, Arg & Pro Mtb	0.28	0.0018	0.047

Table 1: Table of the significant metabolites associated with doxSA levels. Mtb = metabolism.

We then tested for the independence of these association from the doxSA-serine association. By including serine_original as an additional covariate in the model we found that sphinganine_new was again the most associated metabolite (δ=0.36; P(fdr)=0.0014). With sphinganine_new also included in the model, we found six of the ten remaining metabolites belonging to the sub-family of Plasmenilcholines. Four of these were the topmost associated metabolites (result not shown). To test for significance of the entire Plasmenilcholines group, we included the first principal component calculated on this group found a significant negative association (δ=-0.25, p=0.0003). No significant metabolites remained after the additional inclusion of Plasmenilcholines first PC. Exploratory correlation analysis among doxSA, serine_original, sphingnaine_new, and Plasmenilcholines for MacTel cases and controls is presented in **Figure 37**.

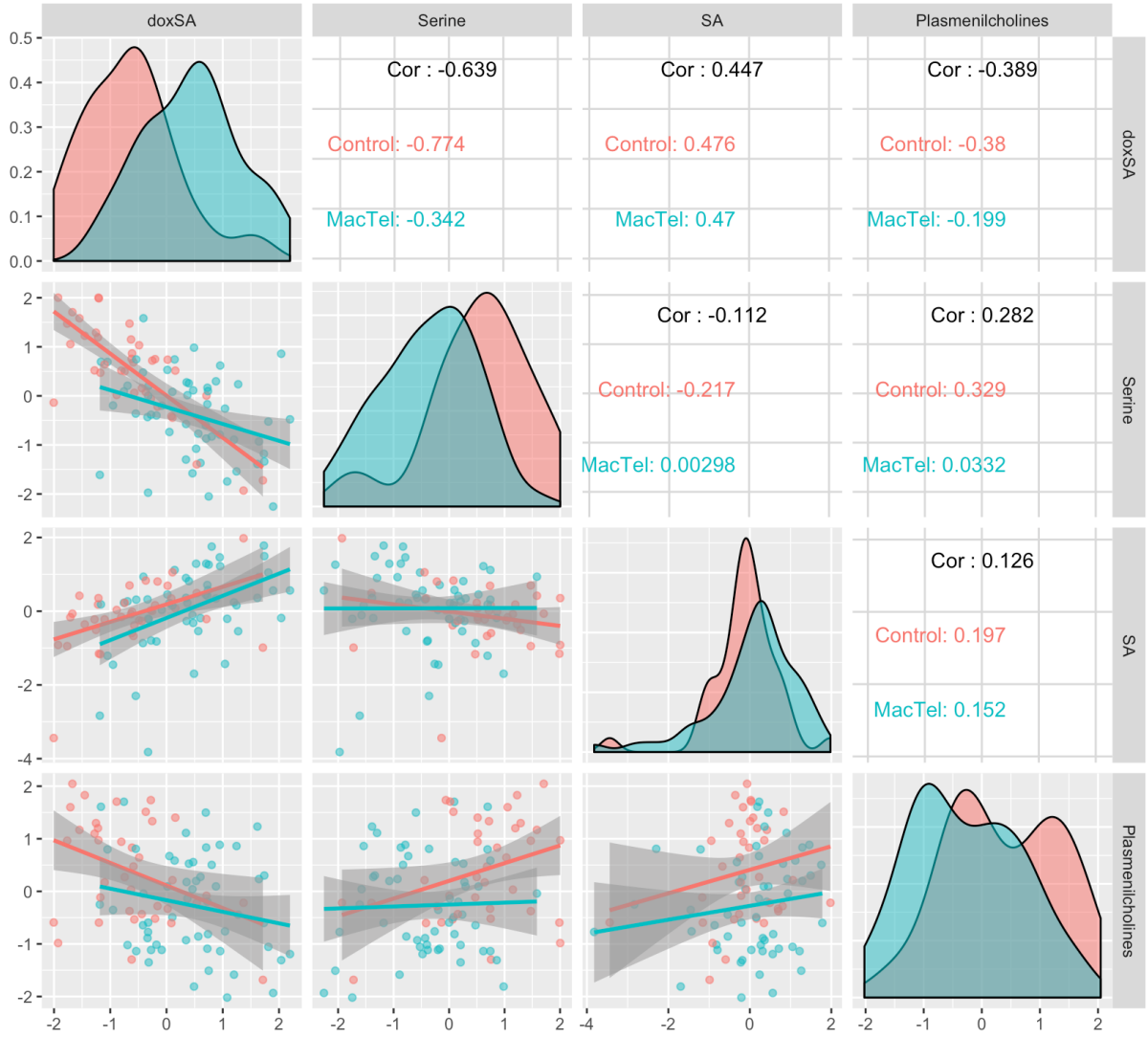


Figure 37: Pearson correlations between doxSA, serine, sphinganine and plamenylcholines (first PC), stratified by MacTel disease status.

As expected from the linear regression model serine, sphinganine and Plasmenilcholines correlated with doxSA but not with each other. Interestingly, almost all correlations were reduced if calculated in MacTel cases only.

Since some of the associations between metabolites and doxSA were observed to be different between MacTel cases and controls we explored which metabolites significantly changed in association with doxSA levels depending on MacTel

status. Although not surviving correction for multiple testing, we found choline to be the metabolite most differentially associated with doxSA depending on MacTel status (δ_3= 0.68, p=1.8e-4, P(fdr)=0.14). In fact, the association between choline and doxSA was completely inverted, at r = -0.45 for controls and r = 0.37 for cases (**Figure 38**).

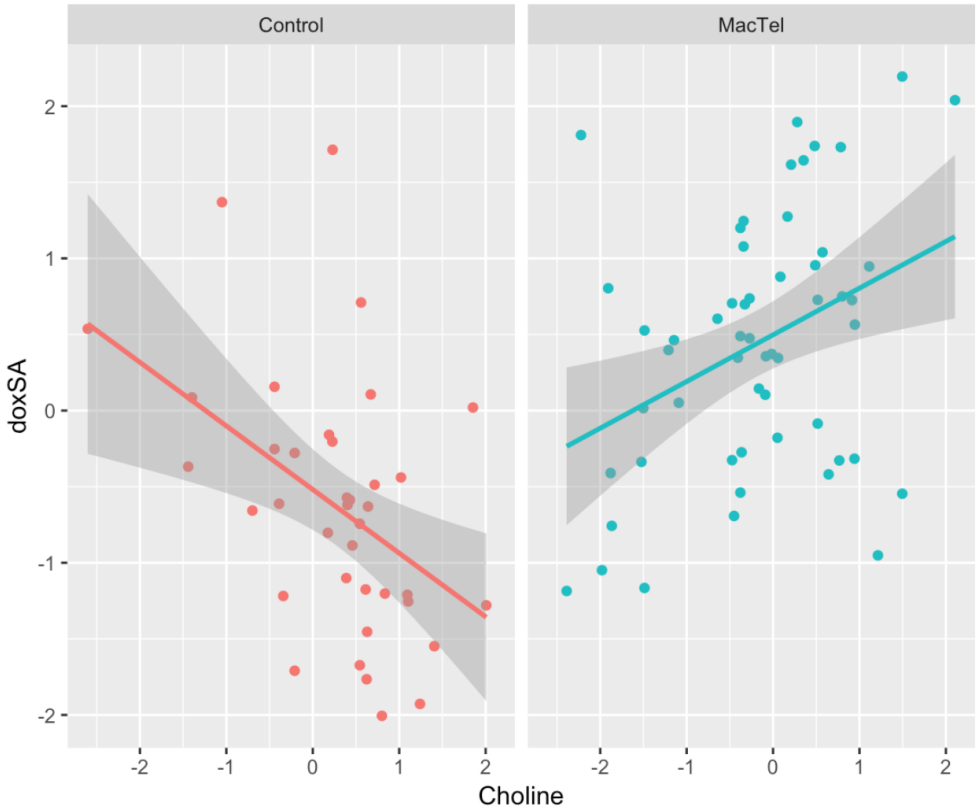


Figure 38: Correlation between doxSA and choline in cases and controls.

We found no significant interaction between any metabolites and doxSA on MacTel disease risk, although, among the top 10 nominal results (0.15<pFDR<0.2) the first 9 were sphingomyelins (1.5<δ_3<1.09, 4.3e-4<p<1.9e-3) and the tenth was serine_original (δ_3=1.38, p=0.0028). Similarly, to the Plasmenilcholines group analysis above we took the first principal component calculated on the sphingomyelin group and tested its interaction with doxSA on MacTel risk. This,

as expected, was nominally significant (δ_3=1.15, p=0.0009). To visually explore the meaning of these interaction terms we divided doxSA levels into three groups, capturing the 1st, 2nd and 3rd tertiles of the doxSA distribution. We then examined the abundance of serine and sphingomyelins first principal components by differentiating between cases and controls in all of the three doxSA abundance groups **Figure 39** (A-B).

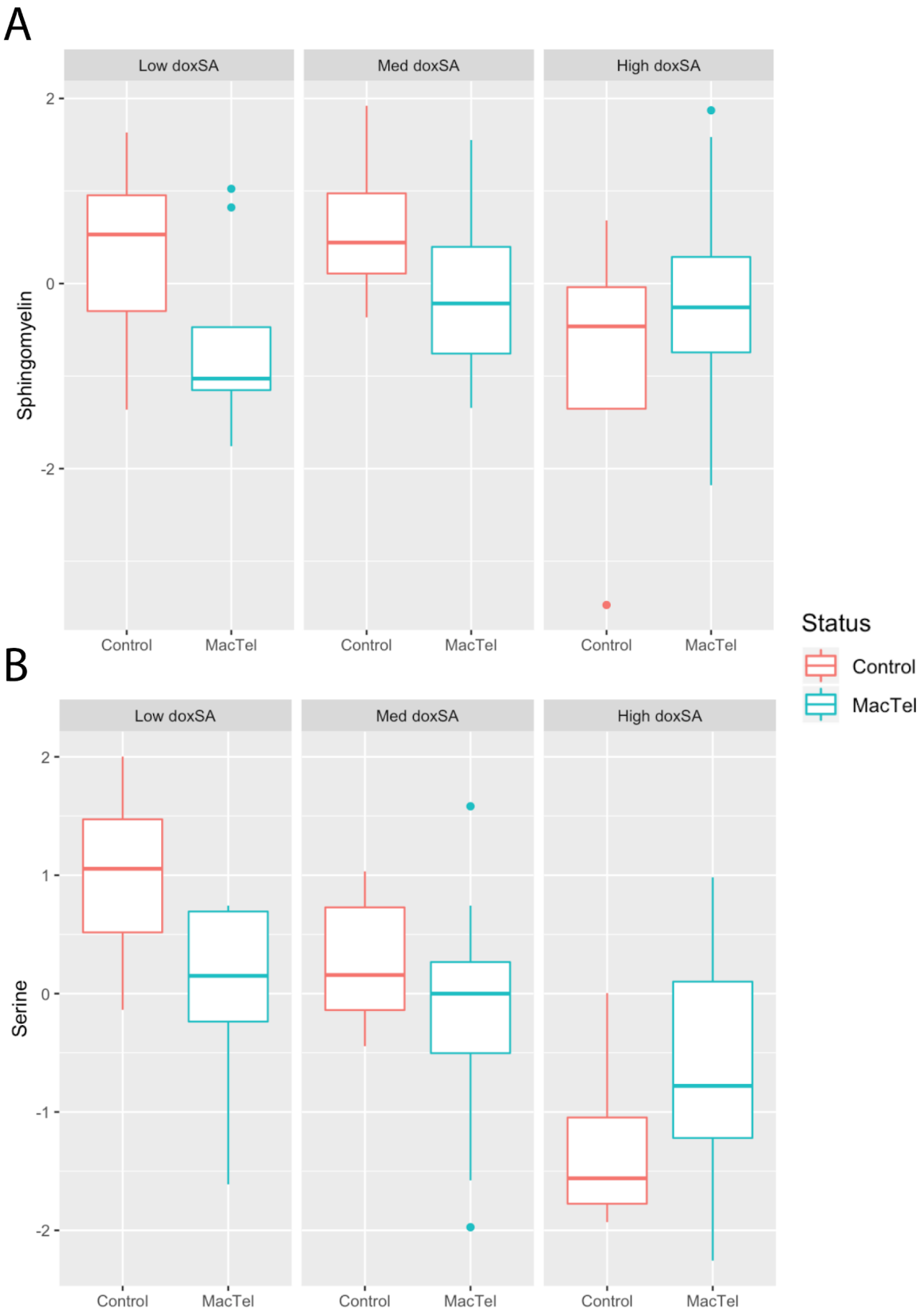


Figure 39: Visualisation of sphingomyelin first principal component abundance (A) and serine abundance (B) stratified by disease status and doxSA tertile.

We observed that sphingomyelin levels were comparable between MacTel cases and controls with high doxSA, but MacTel sphingomyelin abundance decreased much more than controls as doxSA decreases. In contrast, serine was lower than controls in MacTel subjects with low doxSA; and higher than controls in the highest doxSA tertile.

We looked for metabolites still associated with MacTel disease status even after correction for doxSA levels. We found the top two metabolites to be 2-hydroxyglutarate and glycine possibly due to their connection with the gene *CPS1* and partly disconnected from the serine-doxSA-MacTel pathway (results not shown). As a confirmation of this observation of this compartmentalisation of these two metabolites, when including the additional interaction effect between doxSA levels and the first sphingomyelin group principal component, 2-hydroxyglutarate remained significant but the effect of glycine disappeared, likely due to its connection with serine.

When investigating which metabolites might impact the differential association between MacTel and doxSA levels for T2D we interestingly found three plasmenylcholines as the topmost significant metabolites. When including an interaction term between MacTel status, T2D and the Plasmenylcholine first PC we indeed found this to be significant ($p=0.004$). To visually explore these triple interactions we split the Plasmenylcholine first PC by the median value and

compared the doxSA levels among MacTel patients and controls further, stratifying by their T2D status (**Figure 40**).

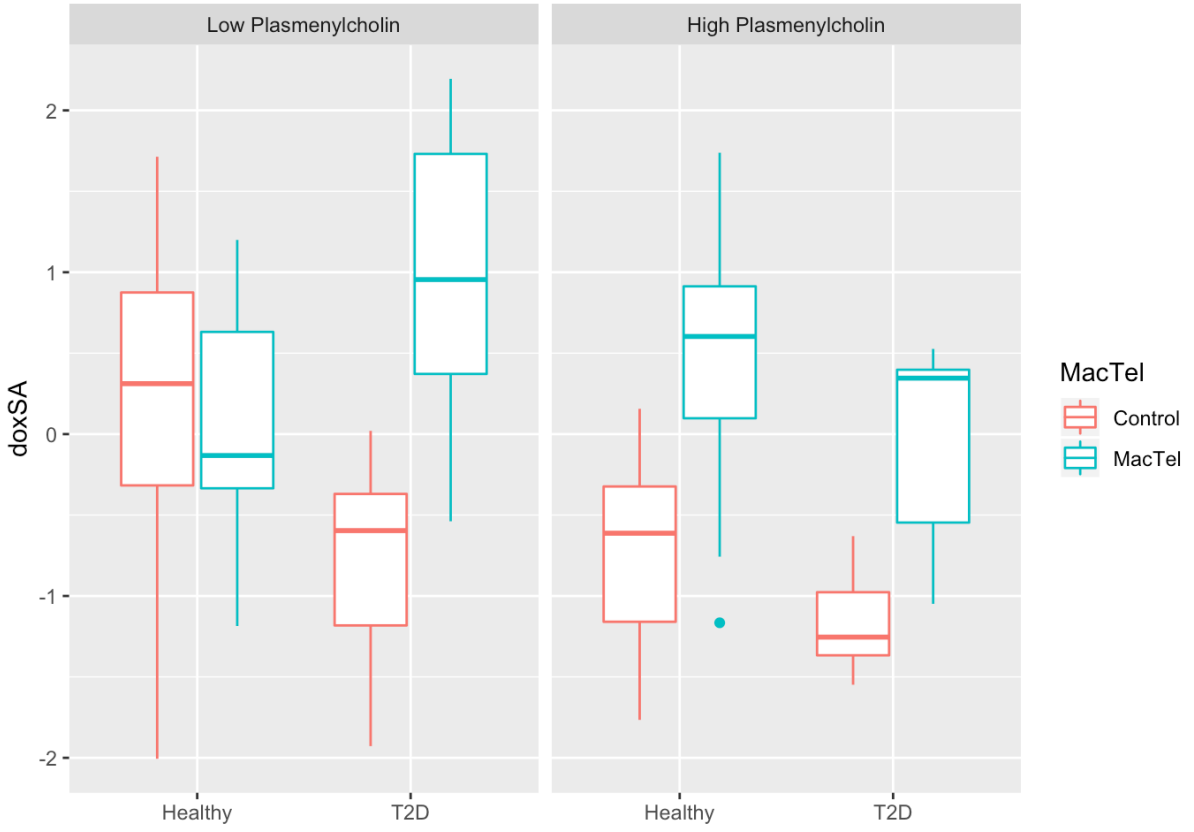


Figure 40: Abundance distribution of doxSA levels divided by MacTel status, T2D status, and plasmenylcholine PC.

This image reveals that in a low Plasmenylcholine environment doxSA levels appear to be different between cases and controls only if the individuals are affected by T2D. Moreover, we see that healthy controls also unaffected by T2D appear to have generally higher doxSA levels compared to the other MacTel

controls. MacTel patients with high plasmenylcholines levels presented higher abundances of doxSA independent of T2D status.

5.4 Discussion

This study investigated connections between the effect of deoxy-sphingolipid accumulation and risk of Macular Telangiectasia Type 2. Through statistical analyses, we further investigated how deoxy-sphingolipid accumulation might arise from either specific rare genetic mutations (eg. *SPTLC1* or *SPTLC2* mutations) or systemic depletion of serine.

5.4.1 MacTel patients with serine depletion accumulate deoxy-sphingolipids in blood and likely in retina

In this study, we observed that mice fed a serine- and glycine-depleted diet accumulated deoxy-sphingolipids in both circulating blood and retinal cells. This is likely to result from SPT utilizing alanine in place of serine. Our previous studies presented in Chapter 2, 3 and 4 detailed depleted levels of serine in MacTel patients, likely caused by specific mutations in genes that modulate serine biosynthesis or glycine availability. Pleasingly, in this study, we observed that MacTel patients have high levels of deoxy-sphingolipids in circulating blood, which correlates negatively with serine and positively with alanine. Concordant with this, alanine was identified in Chapter 3 as a metabolite with a possible causative effect on MacTel risk, where we identified high abundances of genetically predicted alanine as predictive of increased MacTel risk.

Although studies on retinal tissue from individuals affected by MacTel are still needed, in this study lower circulating serine concentrations were sufficient to increase deoxy-sphingolipid levels in the mouse retina. It is reasonable to assume that a similar relationship between human blood and retinal serine concentrations exists. Lastly, it is interesting to note that, as mentioned in Chapter 1, MacTel often occurs in the lower temporal side of the macula and its clinical signs are usually confined to a specific area. Hence, analysis of deoxy-sphingolipids in human retinal samples, including from subjects affected by MacTel, and observing the specific macular layers or regions prone to deoxy-sphingolipid accumulation given the distinct anatomy of this disease might give insights into why MacTel symptoms are confined to the macula in the eye.

5.4.2 Deoxy-sphingolipids are likely not the only MacTel disease drivers

This study found the very first causative genetic mutation that explained the comorbidity of MacTel and HSAN1, suggesting that the accumulation of deoxy-sphingolipids, without serine depletion, is sufficient for the development of MacTel. However, we observed that not all mutations that increase deoxy-sphingolipids and produce HSAN1 were causative for MacTel. In fact, HSAN1 patients with mutation p.Cys133Trp in gene SPTLC1 did not develop MacTel. Moreover, some MacTel patients without HSAN1 mutations were observed to have physiological deoxy-sphingolipids levels comparable to controls. This contrasting finding highlights that deoxy-sphingolipids are not likely to be the only causative

drivers behind MacTel and other metabolic or phenotypic disturbances will need to be taken into account.

5.4.3 Deoxy-sphingolipid accumulation affects retinal health and correlates with MacTel progression

In the mouse diet study, we observed impaired retinal function in serine-deprived animals. As MacTel is associated with abnormalities in retinal cell layers, it is likely that deoxy-sphingolipid accumulation in both blood and retina might be the cause for the insurgence of the first MacTel symptoms. We also noticed that deoxy-sphingolipid abundance suggestively correlated with ellipsoid zone (EZ) loss progression in MacTel patients. As deoxy-sphingolipids are incorporated into cell membranes, and cannot be degraded via beta-oxidation, we expect that MacTel patients slowly accumulate deoxy-sphingolipids over the course of their life. This accumulation might become increasingly toxic for the retina which is likely to deteriorate faster in higher deoxy-sphingolipids surroundings. Interestingly, in Chapter 3 we presented results on the causative role of serine depletion where we found that low levels genetically predicted serine to increase the risk of MacTel progression markers, which further supports an important role for deoxy-sphingolipids accumulation in MacTel progression.

We recognise, however, that the cross-sectional setting of this study was not suited to analyse progression outcomes. Moreover, as mentioned in Chapter 1 we observed in a previous study that EZ loss progression in MacTel patients appears to be non-linear, and is characterised by a slow initial progression with a

subsequent spike in EZ loss, followed by a final plateau, usually confined to the "MacTel area" (16). Additionally, longitudinal measurement of deoxy-sphingolipids in blood will be required to determine their prognostic utility relative to EZ loss. In lieu of this, a suitable proxy for such a study may be measurements of deoxy-sphingolipids in MacTel patients at a single time point combined with longitudinal observations of EZ loss.

5.4.4 Prospects for MacTel treatment: serine supplementation and fenofibrate

We observed that a diet lacking serine and glycine alone might be sufficient to cause deoxy-sphingolipids accumulation which is known to cause MacTel in humans. Oral serine supplementation might, therefore, be used to prevent or limit the accumulation of deoxy-sphingolipids. Glycine supplementation appears not to be a viable treatment as this is thought to be depleted only to counteract serine insufficiency.

High serine abundance has been associated with increased cancer risk and progression (89–91), casting doubt on the promise of serine supplementation in MacTel patients, especially those older than 50 years who are more prone to cancer development. However, we are encouraged by a recent manuscript by Lotta et al which found no causative relationship between serine abundance and cancer risk in a large UK cohort.

As mentioned, not all MacTel patients have depleted serine, and hence elevated deoxy-sphingolipid levels. MacTel patients with HSN1 disease due to mutations affecting the SPT enzyme might be expected to have normal serine levels. Moreover, we observed that alanine increases doxSA levels independently of serine. Hence direct assessment of serine abundance in MacTel patients prior to prescription of serine supplementations would be crucial to evaluate the appropriateness of such treatment as it may lead to unintended outcomes.

The study by Gantner et al also treated serine-deprived mice with compounds altering lipid metabolism. Specifically, they observed that the PPARα agonist fenofibrate was able to rescue cell death in retinal organoids that were previously subjected to elevated levels of deoxy-sphingolipids. The appropriateness of these potential treatment options will depend on the accurate measurement of patient metabolic phenotypes and evaluation of their genetic risk factors, i.e. needs to be personalised.

5.4.5 Sphinganine independently correlates with deoxy-sphingolipids but not MacTel

In addition to results presented in Gantner et al., we found that serine, sphinganine, and plasmenylcholines were each independently associated with higher levels of deoxy-sphingolipids. These independent associations might indicate that the metabolic pathways connecting each metabolite to doxSA are

different. Because of this, it can be speculated that such metabolites might be involved in separate causative disease mechanisms. Our current understanding of sphingolipid metabolism in MacTel (**Figure 37**) involves dependence between serine, and sphinganine, via SPT. However, it is interesting to note that our previous analysis (Chapter 4) and well as the analysis presented here, did not find differential abundances of sphinganine between MacTel patients and controls. In fact, we observed that MacTel patients exhibit comparable levels of sphinganine to healthy individuals, as well as a positive correlation between these molecules and doxSA. To reconcile these results with current biochemical models of sphinganine biosynthesis requires further study, but suggests the existence of a secondary mechanism which compensates the accumulation of doxSA with sphinganine accumulation - a process apparently not related to the disease.

As presented in Chapter 4, untargeted metabolomics analysis identified different metabolites connected to MacTel which are located downstream of the pathway connecting serine to sphinganine and doxSA. In fact, we observe that patients generally have elevated levels of ceramides as well as depleted abundances of sphingomyelins and choline. Understanding these results will require detailed analysis of the sphingolipids pathway, including all sphingolipid subspecies in MacTel patients, in order to elucidate how such imbalances, connect and ultimately affect disease manifestation.

5.4.6 Plasmenylcholine independently affects deoxy-sphingolipid levels and might be key to decode comorbidity between MacTel and type 2 diabetes

We observed that plasmenylcholines negatively affect doxSA levels independent of disease status, and serine and sphinganine serum concentrations. Interestingly, plasmenylcholine glycerophospholipids predicted doxSA variability only after the inclusion of serine and sphinganine in the model. This indicates that several factors might act simultaneously to affect doxSA. To discover unconfounded associations, such factors need to be taken into account at the same time. As presented in Chapter 3, plasmenylcholines were, in concordance with this study, depleted in MacTel patients, further strengthening the assumption of a genuine role of such metabolites on disease development.

Intriguingly, when analysing the effect of different metabolites on the interaction of MacTel status with T2D on doxSA levels, we found plasmenylcholines to be the most significant metabolites. Specifically, we observed that lower plasmenylcholine levels were generally associated with increased doxSA levels, independent of MacTel status, but dependent on T2D status. More specifically, we found that MacTel patients with T2D and low plasmenylcholine levels had some of the highest doxSA levels, while healthy controls with the same characteristics were among the lowest. Little is currently known about the relationship between plasmenylcholines and T2D, however, associations between such lipids and type 1 diabetes have been described previously (92). Our results suggest that

investigations of plasmenylcholines in MacTel patients who also suffer from T2D, compared to those that do not, are required.

5.4.7 Sphingomyelin and serine distinguish MacTel among subjects with low deoxy levels

In this study, we investigated which metabolites might play a role in MacTel disease in the context of different deoxy-sphingolipids strata. With this analysis we found sphingomyelins to best distinguish between cases and controls, depending on doxSA abundances. Specifically, we found that although MacTel patients and healthy individuals with high deoxy-sphingolipids had similar levels of sphingomyelins, this was not true for individuals with low levels of deoxy-sphingolipids where MacTel patients presented substantial depletion of sphingomyelins. This effect is likely to explain some of the heterogeneity that was found in our previous study (Chapter 4) where these lipids were found to be generally depleted in MacTel patients but not to be the most depleted ones.

It may be that sphingomyelins are extremely depleted only among MacTel patients which are affected by a different biological mechanism from doxSA accumulation, making sphingomyelins an important biomarker that needs to be taken into account when investigating individual causes of MacTel cases.

Sphingomyelins are formed downstream of the sphinganine pathway and directly connected with ceramides and glycerophospholipid metabolism, which was found to be dysregulated in MacTel patients (Chapter 4). Moreover, in our previous chapter, sphingomyelins were found to be one of the most significant differentially co-expressed metabolic groups, and the only one connecting the two most differentially abundant metabolomics groups (glycine-serine metabolism and phosphatidylethanolamines).

Further, our conditional GWAS analysis (Chapter 3) revealed loci which contribute to disease aetiology independent of genetic factors that influence the serine-glycine pathway. This analysis identified SNP rs36259, located on gene *CERS4* as an independent contributor to MacTel risk. *CERS4* is a gene whose products are important in the sphinganine and sphingomyelin conversion. *CERS4* products are expected to act independently of doxSA concentrations to influence sphingolipid concentrations. The results presented in this chapter confirm a potential role of sphingomyelin pathway disturbance on MacTel which might contribute to disease aetiology from a separate, but possibly not independent, mechanism of deoxy-sphingolipids accumulation.

5.4.8 Glycine and 2-hydroxyglutarate depletion as additional MacTel biomarkers independent from deoxy-sphingolipids

Lastly, when investigating metabolites associated with MacTel after accounting for deoxy-sphingolipids, glycine and 2-hydroxyglutarate remained significantly depleted in MacTel patients with respect to control, with the latter significant even after inclusion of the doxSA-sphingomyelin interaction effect. These results are expected, since these are involved in processes and pathways other than sphingomyelin biosynthesis.

Levels of 2-hydroxyglutarate are biochemically independent of serine depletion and likely to be influenced by *PHGDH*, located in 1p12, identified as a GWAS hit. Glycine abundance is directly linked to serine abundance. As covered in Chapters 3 and 4 *CPS1,* contained in another GWAS hit, is important for glycine concentrations and not directly involved in serine abundance. In this study, the relationship between glycine and MacTel risk dissipated when we included the significant interaction term between sphingomyelin and doxSA. It can be hypothesised that the additional effect of glycine on disease risk might not be independent of the aforementioned risk brought by sphingomyelin depletion.

The inclusion of doxSA as a covariate in the model may also reveal additional metabolic connections that are easily missed in the preliminary analysis. We suggest therefore that a more careful and in-depth evaluation of the MacTel metabolomic profile, similar to that presented in Chapter 4, should be performed, incorporating the deoxy-sphingolipids to fully assess their contribution to disease risk.

5.5 Conclusions

In this chapter, we investigated the causal role that deoxy-sphingolipids have on MacTel risk and progression. The updated schematic of the study findings is presented in **Figure 41**. This causal relationship is likely to happen through either direct influence of deoxy-sphingolipids biosynthesis or through the indirect effect of serine depletion. However, we demonstrate that this is likely not the only causal mechanism affecting MacTel, and careful evaluation of different metabolic biomarkers are required to inform the likely success of treatments targeting deoxy-sphingolipids. Such considerations are becoming more important in the burgeoning era of personalised medicine. Lastly, we further confirmed the importance of sphingomyelin concentrations in distinguishing MacTel patients with normal deoxy-sphingolipid abundance, and identified the potential role of plasmenylcholine in MacTel and type 2 diabetes comorbidity, which we think warrants further investigation.

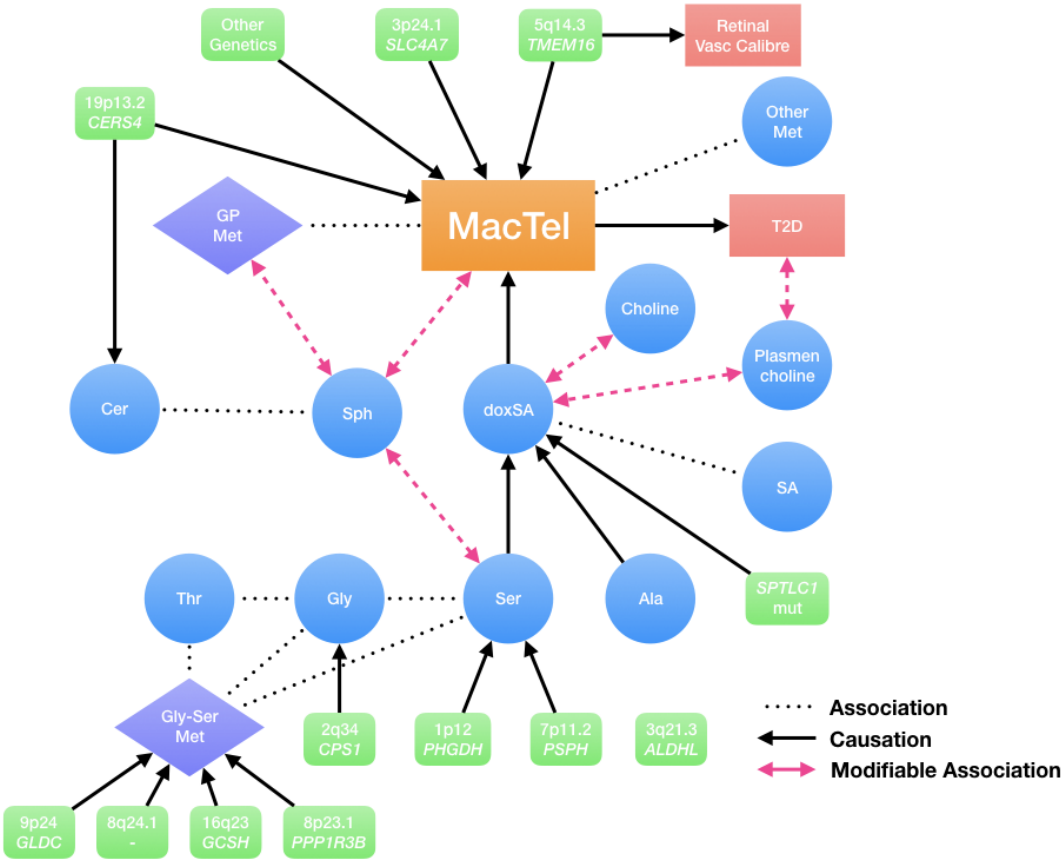


Figure 41: Main findings schematics displaying the drivers and traits associated with MacTel. Genetic traits are displayed as green rounded squares. Metabolites are displayed as blue circles. Metabolic pathways are displayed as purple diamonds. Phenotypic traits are displayed as red squares. Associations are represented by dotted black lines while associations changed by interaction are displayed by dashed red arrows. Causality is indicated by unidirectional solid black arrows.

6 Discussion

In this thesis, we presented four different studies investigating several different aspects of MacTel disease. Each study integrates several types of 'omic data.

The following chapter will summarise and discuss the main findings of these four studies while placing their discoveries on the bigger spectrum of macular and eye disorders as well as suggesting avenues for future studies.

6.1 MacTel is a disease with heterogeneous pathology and genetic causes

One of the main findings from this thesis is that MacTel is indeed a complex disease likely to have multiple causes. In this thesis, we specifically focused on metabolic disturbances involved in MacTel and how this interplay with genetic risks. However, there are still several genetic loci involved in the disease that are independent of the main metabolic dysregulation. This key finding comes from extensive analysis of the data examined in this thesis where these genetic loci remain independent risk factors, even when we correct for all known metabolic drivers and surrogates.

Among the metabolic causes of MacTel we highlight that the accumulation of doxSL reaching toxic levels is likely to be one of the main disease drivers. However,

even within the doxSL framework, we observed a certain level of heterogeneity. We discussed how certain patients might develop a higher abundance of doxSL because of rare mutations on genes catalysing SPT enzymes, impacting the affinity of this enzyme to legate palmitoyl-CoA to alanine rather than serine. This, in turn, results in higher abundances of doxSL even with physiological levels of circulating serine. Other MacTel patients might be presenting with abnormally high levels of alanine, resulting in an imbalance between the two and inducing SPT to use alanine instead of serine. Others again are instead believed to accumulate doxSL because of specific genetic perturbations impairing the biosynthesis of serine. Among these, the genetic mutations impairing serine biosynthesis and its availability are likely to be different. Some patients might be unable to properly produce serine because of genetic impacts on genes such as *PHGDH* and *PSPH*. Others might be unable to convert glycine into serine due to mutations on the *CPS1* gene, resulting in overstimulation of the CPS enzyme which catalyses glycine into other substrates as part of the urea cycle.

With the integration of metabolomics data on MacTel individuals, we discovered many other metabolic disturbances related to the disease. Some of these are likely to be distinct from doxSL. For example, we presented in Chapter 4 and Chapter 5 how sphingomyelin, located downstream from the doxSL biosynthesis, appears to be another driver. Other identified pathways were almost entirely disconnected to doxSL. Among these were phosphatidylethanolamines, lyso-phosphatidylethanolamines, plasmenylcholines, long-chain fatty acids, dyaglycerols, and even xenobiotics.

In addition, we found non-metabolic genetic signatures affecting MacTel risk. The first was locus 5q14.3, tagged by SNP rs73171800, which was associated with both vasculature calibre as well as macular thickness (63, 64, 93, 94). Although MacTel commonly presents symptoms affecting retinal vasculature, we were unable to find any causative effect of vasculature calibre on MacTel disease. The second, locus 3p24.1 (tagged by SNP rs35356316) in chromosome 3 was found by exploring genetic factors affecting disease risk independently from serine genetic influences. This locus is likely to affect the disease by a mechanism which is independent of either metabolic disturbance or retinal vasculature. Interestingly, this locus is within the linkage region for the sensory neuropathy HSAN1b, similar to HSAN1a but additionally presenting cough and gastroesophageal reflux (95). We speculated in the previous chapter that this locus might be involved in altering the expression of the gene *SLC4A4* which would impact the physiological pH of the photoreceptor layers by creating a toxic environment that impedes photoreceptor regeneration. Work on this locus is ongoing.

Heterogeneity was not only observed in the causes of MacTel and their metabolic signature but also in its clinical presentation and progression rate. By combining genomics data with phenomics data, in the form of retinal imaging data, we discovered that MacTel clinical presentation and progression might be modulated by specific genetic influences which in turn might affect different sub-phenotypes.

The final updated schematic on MacTel heterogeneity and some of its possible causes are presented in **Figure 42**.

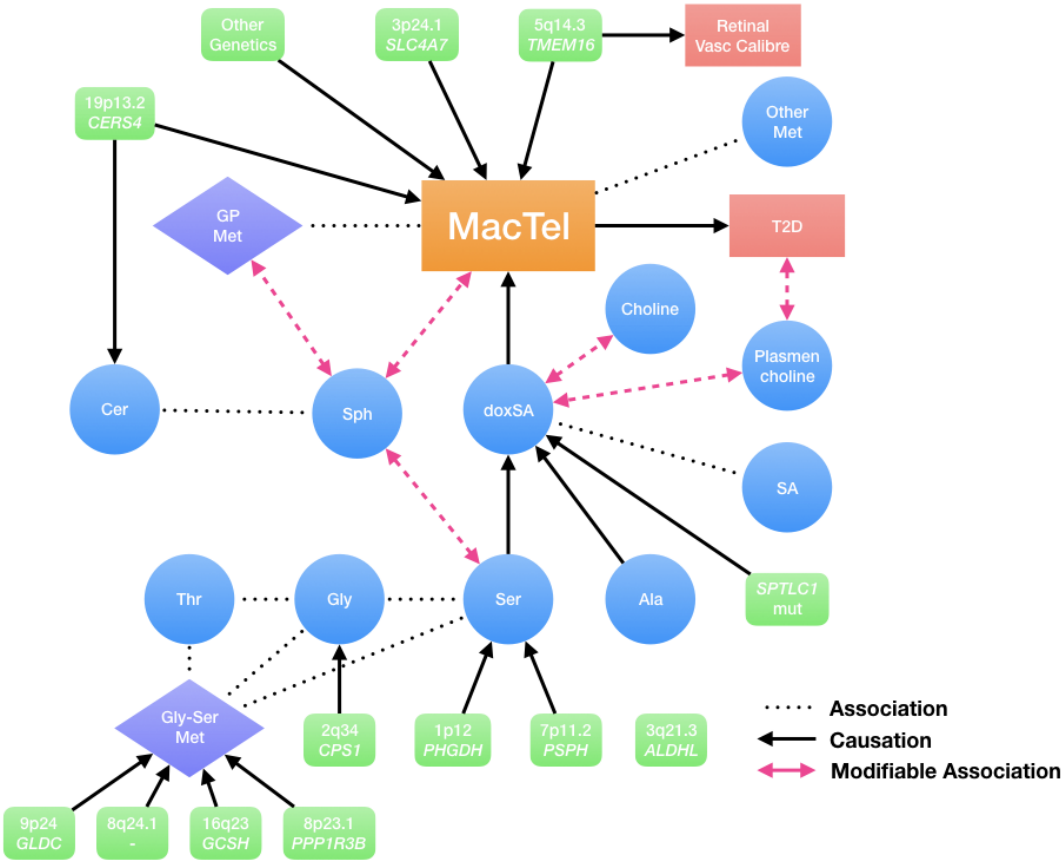


Figure 42: MacTel heterogeneity schematics.

The first GWAS into MacTel already identified multiple genetic loci (61). The work in this thesis has identified further heterogeneity, marking MacTel as an extremely complex diseases, despite its estimated high heritability, at around 70% (Chapter 2). In fact, other more common and likely more heterogeneous disorders have a much lower heritability, such as Age-Related Macular Degeneration (65%) (96) or Diabetic retinopathy (25% - 50%) (97). It is interesting to speculate why MacTel has such high heritability despite this apparent heterogeneity. Several factors are likely to contribute to this. Firstly, the initial GWAS was successful,

identifying four loci, despite its relatively small sample size compared to other GWAS studies performed on more common diseases, including other retinopathies. This indicates that certain loci involved in MacTel had strong enough effect sizes allowing them to be detected and figuratively placing them in the middle of the negative relationship line presented in **Figure 12**. Secondly, we found that even with a moderate sample size of 500 MacTel individuals we were able to cluster loci into functional groups based on their effect on retinal endophenotypes. These results were concordant with external studies. Thirdly, we observed that metabolic signatures of MacTel had a high signal to noise ratio, and even our small-scale metabolomics study (60 cases, 58 controls) was sufficient to identify complex metabolic signatures present in this disease. We even observed correlations of some of these signatures with the progression rate. Lastly, we discovered that there was at least one rare allele affecting the *SPTLC1* gene that is able to cause MacTel by itself, with a very high penetrance, albeit observed with a very low frequency, thereby resolving a small number of the MacTel cases and directly identifying the cause for their pathology as being dysregulation of the sphingolipid pathway. This discovery was possible due to the prior knowledge provided by the GWAS results, resulting in the first Mendelian gene for MacTel.

To conclude, we confirm that MacTel is a complex disorder and further investigations are still required to determine as yet unsolved mechanisms behind this disease pathology. Although complex, we also discuss how observed MacTel heterogeneity is likely to be contained to a much smaller set of pathways. Lastly, we warn that as we move toward the aim of discovering rarer and rarer MacTel

causes, sample sizes required might increase exponentially and careful evaluation of the rarity and uniqueness of such etiological mechanisms need to be assessed based on known disease knowledge.

6.3 Careful evaluation of MacTel patients' metabolic profile is key for treatment prescription

No treatment is yet available for MacTel. Impressive progress has been made with the usage of CNTF implants, which have been shown to slow down or block disease progression (21, 22). However, these implants are invasive, costly and have not been observed to allow recovery of the photoreceptors. Furthermore, their precise mode of action is currently unknown. In this thesis, we discussed which of the factors causing MacTel might be targeted with potential treatments through clinical trials. For example, doxSL levels can be reduced by compounds promoting their catabolism into different subunits, effectively reducing their amount to non-toxicity levels. In fact, as shown in Gantner et al usage of FumonisinB1 and Fenofibrate, which are already known to reduce doxSA, were shown to be sufficient in organoid cell cultures to recover from cell death caused by toxic levels of doxSL. Similarly, serine oral or localised supplementation might be a prospective treatment as increasing serine levels might reduce the production of doxSL and effectively slow the MacTel progression or even allow for cell repairment after doxSL levels have been reduced to normal physiological levels.

Although these considerations are promising for prospective treatment developments, caution is advised. Firstly, we demonstrated and argued that MacTel is not uniquely caused by alteration of the serine to doxSL pathway. For example, certain patients carry the mutation p.Cys133Trp in the *SPTLC1* gene and are expected to present with high abundances of doxSL, despite normal serine levels. For such patients, supplementation of serine would not only be ineffective, but will likely result in a serine overabundance inducing other possible complications. Other patients have normal doxSL levels but severe depletion of serine and sphingomyelin. Similarly, treatment of such patients with either FumonisinB1 or Fenofibrate might prove ultimately ineffective if not deleterious. Through Mendelian Randomization we identified high alanine as a possible additional causal driver for MacTel which might increase levels of doxSL independently of serine abundance. Prospective treatments targeting alanine might be better suited for such individuals. Again, we noticed that plamenylcholine levels could distinguish between MacTel patients with and without T2D. These metabolites, as well as T2D status and insulin levels, should be taken into account when considering prospective treatments. Lastly, we showed how MacTel can develop independently of metabolites and would need a different treatment approach.

To conclude, prospective treatment studies on MacTel using drugs targeting serine or doxSL levels should be a priority. However, careful evaluation and monitoring of the main MacTel biomarkers will be required for patients receiving such treatment to properly evaluate their effectiveness.

6.4 The genetic mechanism affecting vascular phenotypes is still unclear

One of the main clinical signs of patients suffering from MacTel disease is leakage of the retinal vessels on different subretinal layers followed, in certain cases, by abnormal vasculature structure as well as advanced cases of neovascularization. Intriguingly, our initial GWAS identified locus 5q14.3 on chromosome 5, which was previously associated with retinal vascular integrity. However, none of the other loci affecting the vasculature calibre impacted MacTel risk, thereby weakening the initial hypothesis of a causative role of retinal vasculature diameter. Locus 5q14.3s was additionally found in our study and others (94) to impact macular thickness. We recognise that macular thickness can be mostly considered as a bi-standard phenotypes proxy, reflecting a multitude of retinal phenotypes. For example, neuronal degeneration, as well as EZ layer breaks, induce retinal thinning while vascular leakage may induce retinal thickening. This suggests that the effect of locus 5q14.3 might act primarily on retinal vasculature and subsequently resulting in a cascade of factors impacting macular thickness.

In a study by Madelaine et al, the mutations in locus 5q14.3 were found to affect the activity at the CNE1 locus, which acts as an enhancer and cis-regulator of transcription of the microRNA gene called MIR9-2 (64). Suppression of MIR9-2

expression was responsible for several retinal vasculature defects in zebrafish. The same study argued that SNPs variations in the CNE1 locus reduce expression of MIR9-2 predisposing subjects to a lifetime exposure of dysregulated gene activity. This suggests that patients carrying such mutations were predisposed to retinal vascular malformations. The study by Madelaine et al provides a rationale on the possible role of locus 5q14.3 on retinal vasculature phenotypes observed in MacTel patients. As mentioned above, certain MacTel patients not only present vascular leakage and vascular abnormalities but develop neo-vascularisation in later disease stages. This suggests that although mutations in the 5q14.3 locus could explain several vascular malformations in most MacTel patients, neo-vascularization is likely caused by further mutations at different loci. Interestingly, early attempts of VEGF retinal injection - which reduces neo-vascularisation in the wet subtype of Age-Related Macular Degeneration - in MacTel patients were not successful and were deemed deleterious for patients without such phenotypes.

These observations suggest the need for further prospective studies on the role that genetic variants in locus 5q14.3 have on MacTel and especially their role on late-stage neo-vascularisation. Additionally, GWAS could be used to identify different genetic variants that might differentiate between subjects carrying such phenotype from those without and provide further insights into the mechanism responsible for this late-stage clinical sign.

6.5 MacTel might have an inverse causative role on type 2 diabetes

We discussed how the high comorbidity of T2D with MacTel displays an intriguing connection between the two diseases. However, while MacTel patients have a high incidence of T2D, the vast majority of patients with T2D rarely present with MacTel.

In Chapter 3 we found the genetic risk of T2D to increase MacTel risk hence suggesting a potential causative role. However, this association was comparatively small with other causal genetic scores tested in the same study. Diabetes has been connected to glycine depletion and, in our study, we speculated that T2D might not be a causative factor for MacTel disease and we hypothesised that the opposite might be likely. In fact, expecting a causative role of T2D on MacTel should, in theory, translate to most T2D patients presenting with MacTel and only few a few MacTel patients presenting with T2D. Since the opposite is true, we hypothesize that genetic alterations causing metabolic disturbances of MacTel disease are likely to impact on biological mechanisms, like glycine depletion, which in turn influence the incidence of T2D. Intriguingly, when exploring MacTel doxSL levels in Chapter 5, we noticed that diabetic MacTel patients had elevated doxSL levels compared to diabetic controls, while a much lower difference was observed among non-diabetic individuals. We additionally identified plasmenylcholines as key metabolites which could explain the role of doxSL on the comorbidity between

MacTel and T2D. Importantly, MacTel patients usually do not present typical phenotypes of diabetic retinopathy, a retinal disorder affecting some individuals with T2D.

The deep metabolic connections between MacTel and T2D arising from our results still require further investigation and several approaches may be implemented to further explore such relationships. Firstly, a GWAS study restricted to MacTel patients only comparing diabetic and non-diabetic patients is needed to identify distinct genetic signals that might inform on the main causal biological pathways leading to T2D development among MacTel patients. Secondly, a bi-directional MR may prove informative to dissect the genetic relationship between the two pathologies and assess the direction of a causative relationship. However, we highlight that careful evaluation of genetic variants used for such analysis is required and adequate sample sizes are required to robustly divide the two different disease signals. Thirdly, deep metabolomics investigations comparing the metabolic signatures of T2D-MacTel from non-T2D-MacTel may also prove informative. Lastly, correlation of genetic and metabolic signatures of MacTel disease with those of diabetic retinopathy might inform on shared biological influences and enlightening on general retinal health factors.

6.6 Metabolic impact of serine depletion seems to have a broad effect on retinal health

Serine plays a crucial role in MacTel and its depletion results in a clear cascade of metabolic disturbances, ultimately resulting in retinal photoreceptors degeneration. This suggests that other retinal pathologies could be related to serine depletion than those identified so far in the scientific literature. In fact, in work performed by our collaborators (unpublished data), the authors tested for causal associations of genetically predicted metabolites on all disease divided by ICD10 codes on the UK Biobank dataset. Although never significant because of the large multiple testing burden faced in such a study, several eye and retinal pathologies were among the top most significant diseases associated with serine depletion. This suggests that metabolic serine levels might be of preeminent importance for general retinal health suggesting that serine is a key metabolite in the, as yet very poorly understood energy metabolism of the retina. Moreover, we observed that diseases like age-related macular degeneration and diabetic retinopathy are connected with MacTel through shared connections of either clinical signs or common comorbidities. The role of serine on such diseases and more generally in retinal pathologies should be more broadly investigated.

6.7 Future directions

In this section we discuss potential prospective studies that might be performed to either validate or clarify several insights drawn from this thesis. Studies are divided into three main categories (genomics, metabolomics, clinical) and only studies focusing on data analysis and data integration will be presented.

6.7.1 Prospective genomics studies

Our initial GWAS focused entirely on chromosomal SNPs and excluded mitochondrial DNA. A mitochondrial GWAS should be performed to elucidate on the potential role of variants in mitochondrial DNA.

We observed genomic influences on the progression of MacTel. However, analyses performed to check such influences were restrained to datasets never tailored for proper time-related disease progression analysis. A prospective GWAS study and subsequent MR on longitudinal phenotypic MacTel dataset is likely to elucidate the genetic and biological mechanism that might be targeted to slow down or completely stop disease progression.

MacTel patients not only differ from each other in the disease progression rate but also in late-stage neo-vascularisation and T2D comorbidity. When adequate sample size becomes available, GWAS investigations and subsequent MR analysis

on differences between MacTel patients presenting such traits might inform on possible causal mechanisms that further differentiates patients and might instruct for more adequate and specific treatments, deemed to become of pivotal interest in the upcoming era of personalised medicine.

To truly validate an inverse causative role on MacTel to T2D, adequate sample size, as well as better informed instrumental variables used to define MacTel risk, should be used to perform a bidirectional MR comparing T2D with MacTel.

A follow-up study on patients treated with CNTF implants in the light of conclusions drawn in this thesis might help interpret treatment responses and identify patients not likely to derive any beneficial effect from the implants. This, in turn can help plan future trials, targeting those patients most likely to benefit.

The previously mentioned UK Biobank dataset continues to expand and now contains retinal structural images as well as genetic information form tens of thousands of individuals. However, this dataset lacks functional visual assessments which are informative on the visual ability of MacTel patients. Recently a publication predicting fine visual function from structural retinal images in MacTel patients has been published (98), highlighting the possibility to automatically assess functional from structural changes on retinal images. Predicting visual function in such dataset would allow interrogation for causal metabolic influences on general visual impairments using MR, which, as

mentioned earlier, is likely to be impacted by some of the metabolic signatures identified in this thesis.

We found several connections between diabetic retinopathy, age-related macular degeneration and general macular thickness with MacTel. Such traits should then be tested for genetic correlation with MacTel which might inform on the degree of the shared genetic signal between the different traits.

The first MacTel GWAS study attempted disease prediction using SNP data. However, the methodologies used in that study are not comparable to modern machine learning methods that might instead prove extremely useful to investigate a potential prediction tool aiding in clinical diagnosis. Such a tool might become of increasing interest for future population screening studies and for careful diagnosis of MacTel disease and is worthwhile revisiting.

Although we explore eQTLs effect of MacTel candidate SNPs on gene expression, no data was available for retinal eQTLs. Recently a study exploring transcriptional signatures and eQTLs in the human retina has been published (99). This study opens the possibility of assessing retinal transcriptomics influences of MacTel genetic variants as well as performing a Transcriptome Wide Analysis Study, which, through a concept similar to MR, identifies genes which transcription might be causally associated with disease risk.

Lastly, the role of locus 5q14.3 in chromosome 5 and locus 3p24.1 in chromosome 3 is yet to be fully understood, especially in the human retina. Following the informative example published by Madelaine et al, epigenetic studies on such loci in human retinal organoids might elucidate their specific function.

6.7.2 Prospective metabolomics studies

Although MacTel progression is yet to be understood, metabolic dysregulation appears to also play a role. Performing a metabolomics study on longitudinal progression data on MacTel patients has the potential to inform on metabolites which might be targeted for future treatment. Such a study is also necessary to validate our previous findings on doxSL impact on MacTel progression.

A new metabolomics study on MacTel patients also taking into account doxSL levels should be performed. Such a study not only would inform on the broad metabolic impact of doxSL accumulation but could validate the previous finding of the serine and sphingomyelin roles in distinguishing MacTel subjects from healthy controls in a cohort with normal doxSL levels.

6.7.3 Prospective clinical studies

The results presented in the previous chapter outline the need for a prospective clinical trial on the possible treatment of MacTel using serine supplementation in conjunction with doxSL suppressors. However, we again stress how careful

evaluation prior to and during the trial should be exerted on the genetic and metabolic profiles of the enrolled patients. This will ensure a deep understanding of the clinical trials outcomes and treatment efficacy as well as addressing future improvements towards personalised therapies.

On a final note, glycine and serine are available as "over the counter" food supplementation at any pharmacy/chemist. Hence, although a clinical trial is clearly warranted, as outlined above, this will be difficult to perform in the age of consumer awareness and social media. The patients moderated MacTel Facebook Group (954 followers) from August 2017 has posted more than 370 posts informing on all publicly available MacTel studies, studies on other eye disorders as well as supplementations for eye health. The group has indeed not missed the hint of potential involvement of glycine and serine in MacTel and discussions among patients regarding serine and glycine supplementation and diet were already shared right after the Scerri et al 2017 publication. Hence clinical trials will be very difficult to perform in such a climate because compliance, baseline and continued blinding throughout the study may be very difficult to achieve. This is not a challenge unique to MacTel, but MacTel is possibly one of the diseases where therapeutic trials are most compromised going forward.

Appendix A: Scerri et al 2017

Genome-wide analyses identify common variants associated with macular telangiectasia type 2

Thomas S Scerri [1,2], Anna Quaglieri [1,2], Carolyn Cai [3], Jana Zernant[3], Nori Matsunami [4], Lisa Baird [4], Lea Scheppke[5], Roberto Bonelli [1,2], Lawrence A Yannuzzi [3,6], Martin Friedlander [5,7], MacTel Project Consortium [8], Catherine A Egan [9], Marcus Fruttiger [10], Mark Leppert [4], Rando Allikmets [3,11] & Melanie Bahlo [1,2,12]

Idiopathic juxtafoveal retinal telangiectasis type 2 (macular telangiectasia type 2; MacTel) is a rare neurovascular degenerative retinal disease. To identify genetic susceptibility loci for MacTel, we performed a genome-wide association study (GWAS) with 476 cases and 1,733 controls of European ancestry. Genome-wide significant associations ($P < 5 \times 10^{-8}$) were identified at three independent loci (rs73171800 at 5q14.3, $P = 7.74 \times 10^{-17}$; rs715 at 2q34, $P = 9.97 \times 10^{-14}$; rs477992 at 1p12, $P = 2.60 \times 10^{-12}$) and then replicated ($P < 0.01$) in an independent cohort of 172 cases and 1,134 controls. The 5q14.3 locus is known to associate with variation in retinal vascular diameter, and the 2q34 and 1p12 loci have been implicated in the glycine/serine metabolic pathway. We subsequently found significant differences in blood serum levels of glycine ($P = 4.04 \times 10^{-6}$) and serine ($P = 2.48 \times 10^{-4}$) between MacTel cases and controls.

MacTel cases typically present at 40–60 years of age with abnormal right-angled juxtafoveolar capillaries and parafoveal telangiectasias. MacTel is an uncommon disease with a 0.0045–0.1% population prevalence and no obvious sex bias[1–3]. Retinal lesions typically co-present with MacTel, including retinal transparency, outer retinal and choroidal neovascularization, lamellar holes or foveal cysts, photoreceptor dysfunction, minimal exudation, yellow–white parafoveal crystals, and retinal pigment epithelium (RPE) pigmentation abnormalities and atrophy. Central vision impairment and decreased visual acuity are the usual clinical outcomes. MacTel is a bilateral disease, but asymmetry of the eyes for disease severity and presence of lesions is possible. The lesions also occur in 0.06–1.18% of the general population[2].

Risk factors for MacTel are largely unknown; however, associations have been observed with smoking[2,4], diabetes[5,6], high body mass index (BMI)[6], hypertension[6] and obesity[6].

Observations of MacTel-affected monozygotic twins[4,7–9] and multiplex families with vertical transmissions of MacTel[1,5,9–12] suggest a genetic etiology for the disease. The late age of onset, low penetrance and variable phenotype, as exemplified by asymptomatic affected relatives[9], and positive and negative misdiagnoses complicate the discovery of genetic variants predisposing to MacTel. We previously screened 27 candidate genes in eight unrelated MacTel cases but found no causative mutations[13]. Linkage analysis of 17 families including individuals with MacTel identified a 15.3-Mb locus on chromosome 1q41-42.2 (logarithm of odds (LOD) = 3.45); however, sequencing of the underlying genes identified no causative mutations[14].

RESULTS

Discovery GWAS stage

The GWAS discovery stage included genotype data for 6,312,048 SNPs after quality control and imputation (including 1,093,805 SNPs genotyped on Illumina Omni SNP chips) in 476 MacTel cases and 1,733 controls (Table 1 and Online Methods). This sample size was large enough to achieve power of at least 0.90 for risk variants with allele frequencies of 0.10–0.70, assuming a population prevalence of 0.001 for MacTel and an odds ratio (OR) effect size of 2 for the risk allele (Supplementary Fig. 1), but the power drops to 0.53 for an OR of 1.5, as expected from a GWAS of modest size. Samples came from individuals of European descent, as confirmed by comparative analysis with 1000 Genomes Project samples (Supplementary Fig. 2). Population substructure was assessed using principal-component analysis (PCA; Supplementary Figs. 2–4). Heritability on the liability scale for MacTel was estimated as 0.21 or 0.74, given prevalence rates of 0.0045% and 0.1%, respectively. The GWAS was performed with logistic regression modeling including the first principal component (PC1) as a covariate. Quantile–quantile plots (Supplementary Fig. 5) demonstrate the correction of inflation likely due to population

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. [2]Department of Medical Biology, The University of Melbourne, Parkville, Victoria, Australia. [3]Department of Ophthalmology, Columbia University, New York, New York, USA. [4]Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA. [5]The Lowy Medical Research Institute, La Jolla, California, USA. [6]Vitreous Retina Macula Consultants of New York, New York, New York, USA. [7]Department of Cell and Molecular Biology, The Scripps Research Institute, La Jolla, California, USA. [8]A list of members and affiliations appears in Supplementary Table 1 . [9]Medical Retina Department, Moorfields Eye Hospital NHS Foundation Trust, London, UK. [10]UCL Institute of Ophthalmology, University College London, London, UK. [11]Department of Pathology and Cell Biology, Columbia University, New York, New York, USA. [12]Department of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria, Australia. Correspondence should be addressed to M.B. (bahlo@wehi.edu.au).

Received 8 August 2016; accepted 31 January 2017; published online 27 February 2017; doi:10.1038/ng.379 9

Table 1 Sample characteristics

Sample	N_{total}	N_{males}	$N_{females}$	Source [a]	Age range in years [b]	% with diabetes [c]
Discovery cases	476	192	284	MacTel Project	45–86 (mean = 63.2, s.d. = 8.04)	33.4
Discovery controls	1,733	735	998	MacTel Project ($N_{subtotal}$ = 76)	39–84 (mean = 62.2, s.d. = 10.64)	6.6
				AREDS ($N_{subtotal}$ = 1,657)	55–80	6.3
Replication cases	172	74	98	MacTel Project	38–85 (mean = 62.4, s.d. = 8.45)	33.3
Replication controls	1,134	629	505	Columbia University controls ($N_{subtotal}$ = 629)	74.8 ±7.1	–
				CCHMC controls ($N_{subtotal}$ = 505)	2–52	–

[a]Sources included the MacTel Project Consortium, the Age-Related Eye Disease Study (AREDS), Columbia University controls and the Cincinnati Children's Hospital Medical Center (CCHMC). [b]Age extracted from the literature for AREDS participants, Columbia University controls and CCHMC controls. [c]Diabetes status for the Columbia University and CCHMC controls is unknown but presumed to be at baseline levels.

substructure (without PC1, genomic λ = 1.141; with PC1, genomic λ = 1.035). Additional principal components increased the genomic inflation factor with no difference to the key results, likely owing to the small sample size.

SNPs at six loci surpassed the genome-wide significance threshold ($P < 5 \times 10^{-8}$; Fig. 1 and Supplementary Tables 2 and 3). Three of these loci (at 1p36.22, 3p24.1 and 7p21.3) were discounted as false positives after failing technical validation steps (Online Methods). The remaining three loci (at 5q14.3, 2q34 and 1p12) included 149 SNPs that reached genome-wide significance (Fig. 2). Seven of the 149 SNPs (Table 2) were genotyped in an independent, ancestrally European cohort of 172 MacTel cases and 1,134 controls. Selection of SNPs for replication was governed by the ability to create TaqMan assays. We ensured that at least one genotyped, rather than imputed, SNP was selected at each locus. As it was not possible to perform PCA on the replication cohort, the replication analysis was performed without covariates. SNPs at all three loci replicated association ($P < 0.01$; Table 2).

Locus 5q14.3

At 5q14.3, 116 SNPs were genome-wide significant within a 400-kb region, including the strongest associated variant in our study, rs73171800 (per-allele OR = 2.41, 95% confidence interval (CI) = 1.96–2.96; $P = 7.74 \times 10^{-17}$). Two SNPs were selected for replication at this locus, rs17478824 and rs73173548, and both associated with MacTel in our independent cohort (Table 2). Among the other genome-wide significant SNPs at this locus were rs2194025 (per-allele OR = 2.31, 95% CI = 1.88–2.84; $P = 1.97 \times 10^{-15}$) and rs17421627 (per-allele OR = 2.43, 95% CI = 1.93–3.06; $P = 3.51 \times 10^{-14}$). All five of these variants are in strong linkage disequilibrium (LD; pairwise r^2 = 0.73–0.98), and their minor alleles confer risk to MacTel. Similarly, the minor alleles of both rs2194025 and rs17421627 have been associated with increased retinal venular and arterial calibers (Table 3)[15,16]. A broader region encompassing this 5q14.3 locus is implicated in capillary malformation (MIM 163000) and hereditary benign telangiectasia (MIM 187260), two cutaneous vasculature abnormality traits that may be variable expressions of the same disorder with possible within-family comorbidity[17–19]. Hereditary benign telangiectasia has been linked to a 7-Mb region at 5q14.1-14.3 (ref. 17). Capillary malformations have been linked to a 19-Mb region at 5q14.1-15 (refs. 18,20), and germline mutations of the gene RASA1 (encoding RAS p21 protein activator (GTPase-activating protein) 1) were identified as the likely cause for the disorder, albeit with phenotypic variability[21–24]. Murine knockout models for Rasa1 are embryonic lethal and display vascular abnormalities[25]. A microdeletion at 5q14.3 that includes the genes RASA1 and MEF2C (myocyte enhancer factor 2C) is also implicated in capillary malformations[26]. Our top associated SNP rs73171800 is intergenic at this locus, and the broader associated region is within the bounds of the 5q14.3 microdeletion but does not include RASA1 or MEF2C, which are around 1 Mb proximal and 100 kb distal, respectively. The nearest genes are TMEM161B, TMEM161B-AS1, LINC102546226 and LINC00461, which are largely uncharacterized. The role of RASA1 in angiogenesis and capillary malformations makes it an interesting candidate for MacTel. However, MEF2C is closer to the locus, is endothelially expressed and is involved in early murine vascular development[27]. MEF2C haploinsufficiency causes several neurological disorders, including cortical malformations, epilepsy and mental retardation[28,29]. Mef2c is involved in murine synaptic development and regulation of synaptic transmissions[30]. Murine Mef2c knockout models are embryonic lethal[31] and display vascular abnormalities[27], including tissue-dependent lumen size variability and defective vascular remodeling[32]. With respect to the late-onset characteristic of MacTel, initial vasculogenesis in Mef2c-null mice is reported normal, but anomalies arise later on in development, albeit still during embryogenesis[32]. Vascular endothelial growth factor A (VEGFA)[33], a key player in angiogenesis, controls MEF2C expression in retinal epithelial cells. An antiangiogenic role for MEF2C during stress conditions has been suggested[34]. Epithelial cell–specific Mef2c-null mice suppress retinal vascular degeneration and promote retinal vessel regrowth in response to oxygen-induced retinopathy but not during normal, unstressed retinal vasculature development[34]. Mef2c retinal-specific regulation is controlled by an alternative promoter regulated by the transcription factor neural retina leucine zipper (NRL)[35]. Nrl-null mice lack rod function[36], and mutations of NRL cause retinitis pigmentosa 27 (MIM 613750)[37].

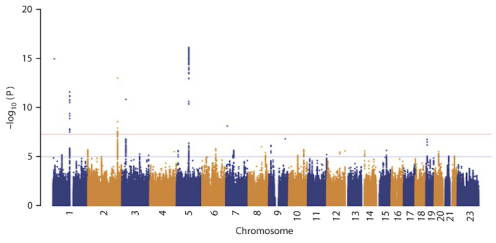


Figure 1 Genome-wide plot of association. SNPs on sequential chromosomes are alternatively colored blue and orange. The x axis represents chromosomal position, and the y axis represents the $-\log_{10}$ (P value) of association for each SNP with MacTel as tested by logistic regression with PC1 as a predictor to correct for population stratification. Analysis was performed with 476 MacTel cases and 1,733 controls. Red and blue horizontal lines correspond to the thresholds for genome-wide significant and suggestive association, respectively.

Locus 2q34

At 2q34, an 83-kb region contained 21 variants reaching genome-wide significance and included the gene CPS1 (carbamoyl-phosphate synthase 1, mitochondrial). This gene encodes a rate-limiting mitochondrial enzyme that performs the first step of the urea cycle, converting

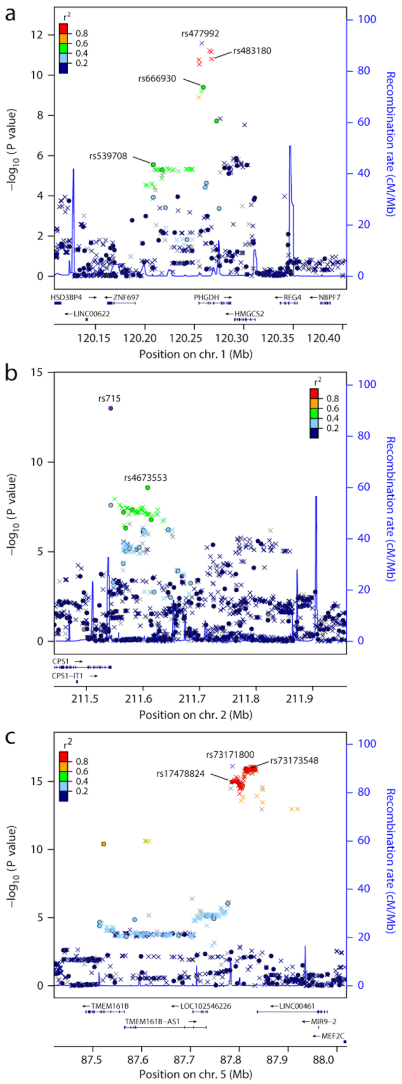


Figure 2 Regional plots of association. (a–c) Association results are shown for the analyzed SNPs, along with recombination rates (blue lines) and the location of known genes (labeled blue horizontal lines), within the three loci associated with genome-wide significance at 1p12 (a), 2q34 (b) and 5q14.3 (c). For each plot, the x axis is scaled such that 100 kb flanks each end of the locus as defined by the first and last SNPs reaching suggestive significance ($P < 1 \times 10^{-5}$), and the y axis represents the $-\log_{10}$ (P value) of association for each SNP. Individual SNPs are represented as colored circles (genotyped) or crosses (imputed). For a–c, the sole purple circle or cross represents the most significantly associated SNP at that locus in the discovery stage and the other markers are colored according to their LD with that SNP. Recombination rates and LD were estimated from the EUR (European) 1000 Genomes Project March 2012 release, build hg19.

ammonia and hydrogen carbonate into carbamoyl phosphate (Fig. 3). Mutations of CPS1 cause carbamoylphosphate synthetase I deficiency (MIM 237300) and may lead to death by hyperammonemia[38]. Our most significant SNP at this locus, rs715 (per-allele OR = 0.50, 95% CI = 0.42–0.60; $P = 9.972 \times 10^{-14}$), resides within the 3′ UTR of CPS1. We replicated associations with rs715 and rs4673553, another genome-wide significant SNP at this locus, in our independent cohort (Table 2). We found that the major allele of rs715 confers risk for MacTel. Other GWAS have associated rs715 with several blood metabolite levels[39–42]: in particular, the major allele was associated with decreased glycine[39,40,42–44] and serine[42] (Table 3). Our estimates for the rs715 effect sizes for MacTel are comparable to those for glycine levels[42,43] (Supplementary Fig. 6). Another SNP at this locus, rs2216405, has also been associated with blood plasma glycine levels[45,46] and metabolic ratios involving glycine[47], and it approached genome-wide significance in our study (per-allele OR = 0.54, 95% CI = 0.43–0.68; $P = 9.426 \times 10^{-8}$). These three SNPs are in moderate LD (pairwise $r^2 = 0.27–0.48$). The variant rs1047891 (renamed from rs7422339) causes a threonine-to-asparagine missense substitution within CPS1 and has been associated with blood plasma levels of glycine[48,49], other metabolites[48,50–58], blood flow[56] and vasodilator response[56]. In our study, rs1047891 was imputed; it was excluded because of a low imputed genotype call rate (93.4%) but otherwise would have been significant with the remaining available data ($P = 1.784 \times 10^{-11}$). There is strong LD between rs715 and rs1047891 (pairwise $r^2 = 0.97$). Several studies report a marked sex effect for CPS1 associations, with females showing stronger association for rs715 with glycine levels[43,44] and rs1047891 with homocysteine levels[54,55]. We also observed a sex interaction with MacTel at this locus (Online Methods and Supplementary Table 4). We found that females are at nearly two times greater risk of MacTel for each additional copy of the rs715 risk allele (female-specific per-risk-allele OR = 2.58, 95% CI = 2.00–3.36; male-specific per-risk-allele OR = 1.40, 95% CI = 1.08–1.81). Our estimates for the rs715 sex-specific effect sizes for MacTel are comparable to those for glycine levels[43] (Supplementary Fig. 6).

Locus 1p12

At 1p12, a region of 46.4 kb contained 12 genome-wide significant SNPs. We replicated this signal in our independent cohort with the imputed SNP rs483180 (Table 2). This region includes the genes PHGDH (phosphoglycerate dehydrogenase) and HMGCS2 (3-hydroxy-3-methylglutaryl-CoA synthase 2, mitochondrial). GWAS have associated blood plasma serine levels with intronic PHGDH variants[43,45], including our most significant discovery-stage SNP at this locus, rs477992 (per-allele OR = 1.70, 95% CI = 1.47–1.97; $P = 2.60 \times 10^{-12}$), and also rs478093 (per-allele OR = 1.67, 95% CI = 1.44–1.94; $P = 2.89 \times 10^{-11}$). Here we found that the minor alleles of rs477992 and rs478093 confer risk for MacTel, and the same alleles are associated with decreased serine levels in other studies (Table 3)[43,45]. PHGDH encodes a rate-limiting enzyme that performs the first step of the phosphorylated pathway of serine biosynthesis, converting 3-phosphoglycerate into phosphohydroxy pyruvate (Fig. 3). PHGDH mutations may lead to PHGDH deficiency (MIM 601815), a severe neuropathological disorder with reduced plasma and cerebrospinal fluid serine levels[59]. HMGCS2 encodes mitochondrial 3-hydroxy-3-methylglutaryl-CoA synthase (HMG-CoA synthase-2), an enzyme integral to ketogenesis. Mutations of HMGC2 lead to hypoketotic hypoglycemia due to HMG-CoA synthase-2 deficiency[60].

Suggestive loci

There were 25 loci that showed suggestive evidence for association with MacTel ($1 \times 10^{-5} > P \geq 5 \times 10^{-8}$; Supplementary Table 2). Given

Table 2 Significant associations of three validated and two suggestive loci with MacTel in the discovery and replication stages

Locus	Most significant discovery-stage SNP at locus (imputed?)	OR (95% CI) P value[c]	Gene(s) (genetic location)	SNPs selected for replication: SNP selected for replication (imputed?)	Position (r^2 LD to most significant SNP at locus)[a]	1000G minor allele (freq.)[b]	Stage	Minor allele frequency: Cases	Minor allele frequency: Controls	OR (95% CI) per minor allele[c]	P value[c]
Validated loci											
1p12	rs477992 (yes)	1.70 (1.47–1.97) $P = 2.60 \times 10^{-12}$	PHGDH (intronic)	rs539708 (no)	120,208,503 (0.56)	T (0.415)	Discovery	0.4853	0.3904	1.42 (1.22–1.64)	2.82×10^{-6}
							Replication	0.4394	0.3919	1.23 (0.91–1.67)	0.1761
							Meta-analysis	0.4774	0.3910	1.38 (1.21–1.57)	1.54×10^{-6}
				rs666930 (no)	120,258,970 (0.53)	T (0.458)	Discovery	0.5830	0.4602	1.60 (1.38–1.85)	4.02×10^{-10}
							Replication	0.4949	0.4623	1.14 (0.85–1.54)	0.3744
							Meta-analysis[d]	0.5679	0.4610	1.39 (1.00–1.92)	0.047
				rs483180 (yes)	120,267,505 (0.98)	G (0.315)	Discovery	0.4258	0.3021	1.68 (1.44–1.95)	1.59×10^{-11}
							Replication	0.4142	0.3115	1.64 (1.26–2.13)	0.0002543
							Meta-analysis	0.4227	0.3046	1.67 (1.46–1.90)	1.76×10^{-14}
2q34	rs715 (no)	0.50 (0.42–0.60) $P = 9.97 \times 10^{-14}$	CPS1 (3′ UTR)	rs715 (no)	211,543,055 (1.00)	C (0.291)	Discovery	0.1796	0.3090	0.50 (0.42–0.60)	9.97×10^{-14}
							Replication	0.2135	0.3213	0.58 (0.40–0.82)	0.002384
							Meta-analysis	0.1853	0.3138	0.52 (0.44–0.61)	1.14×10^{-15}
				rs4673553 (no)	211,608,379 (0.41)	G (0.458)	Discovery	0.3466	0.4570	0.63 (0.55–0.74)	2.66×10^{-9}
							Replication	0.3505	0.4641	0.62 (0.45–0.84)	0.002378
							Meta-analysis	0.3473	0.4598	0.63 (0.55–0.72)	2.38×10^{-11}
5q14.3	rs73171800 (yes)	2.41 (1.96–2.96) $P = 7.74 \times 10^{-17}$	TMEM161B – LINC00461 (intergenic)	rs17478824 (no)	87,785,624 (0.97)	T (0.086)	Discovery	0.1775	0.0866	2.34 (1.90–2.87)	1.08×10^{-15}
							Replication	0.2121	0.0855	2.94 (2.00–4.33)	3.90×10^{-8}
							Meta-analysis[d]	0.1835	0.0861	2.47 (2.03–300)	1.09×10^{-19}
				rs73173548 (no)	87,813,971 (0.89)	G (0.080)	Discovery	0.1712	0.0796	2.46 (1.99–3.04)	1.16×10^{-16}
							Replication	0.1667	0.0824	2.21 (1.48–3.32)	0.0001199
							Meta-analysis	0.1704	0.0807	2.40 (1.99–2.90)	7.27×10^{-20}
Suggestive loci											
3q21.3	rs9820286 (yes)	0.61 (0.49–0.75) $P = 5.38 \times 10^{-6}$	ALDH1L1–KLF15 (intergenic)	rs9880406 (no)	126,049,305 (0.88)	A (0.199)	Discovery	0.1345	0.1982	0.63 (0.52–0.77)	8.98×10^{-6}
							Replication	0.1294	0.1898	0.63 (0.45–0.88)	0.00708
							Meta-analysis	0.1331	0.1949	0.63 (0.53–0.75)	2.07×10^{-7}
7p11.2	rs4948102 (yes)	1.46 (1.25–1.70) $P = 2.25 \times 10^{-6}$	PSPH (intronic)	rs4535700 (no)	56,045,448 (0.94)	T (0.258)	Discovery	0.3319	0.2458	1.43 (1.22–1.67)	7.55×10^{-6}
							Replication	0.3500	0.2687	1.44 (1.14–1.83)	0.002477
							Meta-analysis	0.3367	0.2548	1.43 (1.26–1.63)	6.54×10^{-8}
				rs11238389 (yes)	56,079,744 (1.00)	A (0.273)	Discovery	0.3398	0.2525	1.43 (1.23–1.68)	7.04×10^{-6}
							Replication	0.3550	0.2992	1.28 (0.99–1.64)	0.05517
							Meta-analysis	0.3439	0.2649	1.39 (1.21–1.58)	1.42×10^{-6}

[a]Position in base pairs based on the hg19 reference genome and r^2 LD estimated from the discovery cohort sample genotypes. [b]Minor allele and frequency determined from the 1000 Genomes Project (1000G) European population. [c]Odds ratios and P values were derived from a logistic regression model, using PC1 as a covariate in the discovery stage. [d]While all meta-analyses shown were performed with random-effects modeling, a difference from fixed-effects modeling was observed with only two SNPs.

Table 3 Direction of association of MacTel risk alleles with respect to metabolite levels, cis eQTLs and other traits for the three validated and two suggestive loci

MacTel risk locus	Nearest gene(s)	Direction of effect with respect to MacTel risk haplotypes[a]	
		Increased	Decreased
Validated loci			
1p12	PHGDH	eQTL: ZNF697 (2)	Serine
		eQTL: PHGDH (1)	eQTL: PHGDH (1)
2q34	CPS1	Fibrinogen	Glycine
		Betaine	Creatine
		Pyroglutamine	N-acetylglycine
		Glutaroyl carnitine	X-08988
		Homoarginine	Serine
		Nitric oxide metabolites	Glycine/threonine ratio
		Vasodilator response	
		Nitropusside-induced bloodflow	Homocysteine
		High-density lipoprotein	
5q14.3	TMEM161B – LINC00461	Retinal venular caliber	None observed
		Retinal arterial caliber	
		eQTL: TMEM161B -AS1 (20)	
Suggestive loci			
3q21.3 [b]	ALDH1L1 – KLF15	Clamp-based insulin sensitivity [a]	None observed
		Glycine/serine ratio [b]	
		Delta-POST [c]	
		Incident ischemic stroke risk	
7p11.2	PSPH	Delta-POST [c]	Serine
		eQTL: CCT6A (3)	eQTL: SUMF2 (5)
		eQTL: GBAS (6)	
		eQTL: NUPR1L (1)	
		eQTL: PSPH (1)	

[a]Direction of effect for cis eQTLs (observed in the given number of tissues) as derived from the GTEx Project. Metabolite levels are as observed in blood plasma or serum. [b]At 3q21.3, the minor alleles of rs1107366 and rs10934753 were marginally associated with increased risk for MacTel. [c]Difference between pre- and post-methionine load test for circulating homocysteine levels.

our associations with CPS1 and PHGDH, two genes implicated in serine and glycine blood plasma levels, we explored our 25 suggestive loci for overlap with GWAS for these metabolite levels[39,40,42,43,45] and identified the loci at 3q21.3 and 7p11.2 (Fig. 4).

At 3q21.3, rs1107366 has been associated with blood plasma glycine to serine ratios and clamp-based insulin sensitivity[43], and rs10934753 has been associated with homocysteine levels and incident ischemic stroke[58]. Here these two specific SNPs were not associated with MacTel (rs1107366, P = 0.021; rs10934753, P = 0.072); however, two other SNPs at this locus were suggestively associated, rs9820286 (per-allele OR = 0.61, 95% CI = 0.49–0.75; $P = 5.38 \times 10^{-6}$) and rs9880406 (per-allele OR = 0.63, 95% CI = 0.52–0.77; $P = 8.98 \times 10^{-6}$). While rs1107366 and rs10934753 are in strong LD ($r^2 = 0.86$) and so are rs9820286 and rs9880406 ($r^2 = 0.88$), the LD between these two pairs is very low ($r^2 < 0.012$), suggesting that they are tagging two separate signals. We were able to replicate the association with MacTel for rs9880406 in our independent replication cohort (Table 2). This locus lies between the genes ALDH1L1 (aldehyde dehydrogenase 1 family, member L1) and KLF15 (Kruppel-like factor 15). ALDH1L1 is a well-characterized gene that encodes the cytosolic enzyme 10-formyltetrahydrofolate dehydrogenase, which catalyzes the conversion of 10-formyltetrahydrofolate into tetrahydrofolate and carbon dioxide or formic acid (Fig. 3)[61]. ALDH1L1 is a highly specific antigenic marker of astrocytes[62], a type of glial cell critical for retinal angiogenesis during development[63–65]. Transcriptome analysis suggests that astrocytes have an enrichment of metabolic pathways for amino acids such as serine, glycine and cysteine and pathways for the production and processing of glutamate into glutamine for transport to neurons[62]. Mutations of another aldehyde dehydrogenase, ALDH3A2, may cause Sjögren–Larsson syndrome (MIM 270200), a multifaceted neurological disorder with juvenile macular dystrophy[66]. Unlike other maculopathies, MacTel and Sjögren–Larsson syndrome have co-occurring blue light reflectance abnormalities and retinal crystalline deposits. KLF15 encodes a transcription factor that regulates the gene LRP5 (low-density lipoprotein receptor–related protein 5), mutations of which have been associated with the dysplastic retinal vasculature disorder exudative vitreoretinopathy 4 (MIM 603506)[67].

At 7p11.2, rs4947534 and rs4948102 are associated with blood plasma serine levels[42] and homocysteine levels[58], respectively. We found suggestive genome-wide significance for both rs4947534 (per-allele OR = 1.44, 95% CI = 1.23–1.68; $P = 5.38 \times 10^{-6}$) and rs4948102 (per-allele OR = 1.46, 95% CI = 1.25–1.70; $P = 2.25 \times 10^{-6}$) in our discovery stage. Furthermore, our strongest signal at this locus was for rs4948102. We selected SNPs rs4535700 and rs11238389 for replication testing, and rs4535700 was nominally significant (P < 0.01). The variants rs4947534, rs4948102 and rs11238389 are in perfect LD ($r^2 = 1$) and are in high LD with rs4535700 ($r^2 > 0.94$). The minor allele of these SNPs is associated with MacTel risk, and, similarly, the minor allele of rs4947534 is associated with reduced serine levels[42] and increased homocysteine levels[58] (Table 3). Both rs4947534 and rs4948102 are within the gene PSPH (phosphoserine phosphatase), the protein product of which catalyzes the final step in the synthesis of serine by converting phosphoserine into serine (Fig. 3). PSPH mutations may cause phosphoserine phosphatase deficiency, a syndrome with multiple clinical features including reduced plasma levels of serine[68].

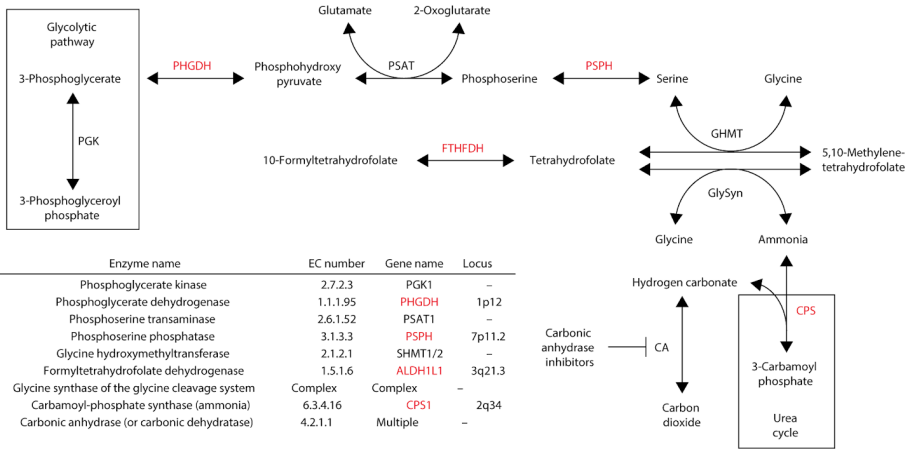
These suggestive loci require validation in independent cohorts to confirm association with MacTel.

Prediction modeling

Our final model incorporated the three validated and two suggestive loci with additive effects, PC1 and one interaction term between sex and 2q34 (Online Methods). Additive modeling was applied because univariate logistic regression analysis showed that it had no significant improvement at these five loci when compared to dominant or recessive models (Supplementary Table 5). As the inclusion of PC1 made only a nominal improvement to the prediction model in the discovery cohort (Supplementary Table 6) and because we could not derive PC1 for the replication cohort, we estimated the probability of MacTel for each sample in the training set (discovery cohort; Supplementary Figs. 7–9) and validation set (replication cohort; area under the curve (AUC) = 0.679; Supplementary Figs. 10–12) without PC1. We produced receiver operating characteristic (ROC) curves (Supplementary Figs. 13 and 14) and computed positive predictive values (PPVs) and negative predictive values (NPVs) (Supplementary Figs. 15 and 16, and Supplementary Table 7) on the basis of these predictions.

Metabolomics analysis

Blood serum levels of 799 metabolites, including glycine and serine, were compared between 50 MacTel cases and 50 independent controls matched for age, sex, ancestry and diabetes status (Online Methods). Student's t-tests were significant for both glycine (T = −6.98; $P = 3.51 \times 10^{-10}$) and serine (T = −5.39; $P = 4.79 \times 10^{-7}$). Given an observed inflation in our results (Supplementary Fig. 17), we applied a correction akin to 'genomic control' with an inflation factor of 1.753 ($P_{glycine} = 4.04 \times 10^{-6}$; $P_{serine} = 2.48 \times 10^{-4}$). Serum levels were lower in the MacTel cases than in the controls (glycine, $\log_2$ (fold change) = −0.541;



Reaction ID	Enzyme name	EC number	Gene name	Locus
PGK	Phosphoglycerate kinase	2.7.2.3	PGK1	–
PHGDH	Phosphoglycerate dehydrogenase	1.1.1.95	PHGDH	1p12
PSAT	Phosphoserine transaminase	2.6.1.52	PSAT1	–
PSPH	Phosphoserine phosphatase	3.1.3.3	PSPH	7p11.2
GHMT	Glycine hydroxymethyltransferase	2.1.2.1	SHMT1/2	–
FTHFDH	Formyltetrahydrofolate dehydrogenase	1.5.1.6	ALDH1L1	3q21.3
GlySyn	Glycine synthase of the glycine cleavage system	Complex	Complex	–
CPS	Carbamoyl-phosphate synthase (ammonia)	6.3.4.16	CPS1	2q34
CA	Carbonic anhydrase (or carbonic dehydratase)	4.2.1.1	Multiple	–

Figure 3 Pathway analysis. Indicated are the enzymes encoded by genes associated with MacTel on chromosomes 1–3 and 7 (highlighted in red). A subsection of the glycolytic pathway highlights the enzyme encoded by PGK, whose differential expression has been observed in MacTel. The enzymes encoded by PHGDH and PGK share the common substrate 3-phosphoglycerate. Also shown is the position that carbonic anhydrase inhibitors, used to treat MacTel, may have in this pathway. For simplicity, common substrates (such as protons, water, NADH, NADPH, ATP and their derivatives) are not shown. Adapted from the ExPASy Bioinformatics Resource Portal [84]. EC, Enzyme Commission.

serine, $\log_2$ (fold change) = −0.382). Of all 799 metabolites, glycine and serine were ranked first and third most significant, respectively. Another glycine/serine pathway amino acid, namely threonine, was the second most significant metabolite, thereby reinforcing our glycine and serine results. Sample genotype data were not available.

In silico variant functional exploration

We used the online GTEx Project[69] portal to search for expression quantitative trait locus (eQTL) signatures in the SNPs showing at least suggestive association ($P < 1 \times 10^{-5}$) at each of the three validated and two suggestive loci (Table 3, Online Methods, Supplementary Figs. 18–22 and Supplementary Tables 8 and 9). The SNPs at 2q34 and 3q21.3 did not yield any cis eQTLs.

At 1p12, SNPs with the strongest association to MacTel also showed the strongest association to gene expression levels, with eQTLs observed in two tissues each for both ZNF697 and PHGDH. For PHGDH, the direction of effect was tissue specific.

At 5q14.3, MacTel risk alleles associated with increased expression of TMEM161B-AS1, the antisense RNA to TMEM161B, in nearly every tissue tested. However, the genome-wide significant 5q14.3 MacTel risk SNPs ($P < 5 \times 10^{-8}$) were not driving the majority of these eQTL associations; instead, the eQTLs were driven by the SNPs showing suggestive association with MacTel ($1 \times 10^{-5} > P \geq 5 \times 10^{-8}$). This suggests that, while the MacTel risk SNPs are indeed affecting the expression of TMEM161B-AS1, they may be having a stronger unobserved functional effect on TMEM161B-AS1 or other genes.

eQTLs at the suggestive 7p11.2 locus were observed in multiple genes for a limited number of tissues, and the direction of effect was gene specific. Similar to the 1p12 locus, the SNPs with the greatest association to MacTel risk also had the greatest association to gene expression levels.

Methylation and H3K4me3 histone modification architecture around the three significant and two suggestive loci were analyzed in the UCSC Genome Browser (Supplementary Figs. 18–22). H3K4me3 was selected because of its known association with transcriptional activity. At the 1p12 locus, there is a strong H3K4me3 signature and a moderate unmethylated DNA signature around the strongest associated MacTel SNPs at the start of PHGDH. This supports the eQTL results and suggests that these SNPs may affect expression of PHGDH. At 2q34 and the suggestive 3q21.3 locus, there are no notable H3K4me3 signatures, but a weak methylation signature is evident at 2q34, indicative of gene silencing. At 5q14.3, there is a small but robust H3K4me3 mark at the most strongly associated MacTel risk SNPs downstream of TMEM161B-AS1. Finally, at 7p11.2, the suggestive MacTel risk SNPs do not overlay any H3K4me3 modification or methylation signatures, but they do lie immediately adjacent to such signatures, suggesting that they could be influencing the DNA structure by virtue of their proximity.

Interpretation of these results is difficult given that we do not have any strong a priori expectation of the particular tissues that would be of interest for MacTel and the absence of eye tissue in the GTEx resource.

DISCUSSION

Here we identified three validated and two suggestive genetic loci for MacTel. While these loci include several plausible candidate genes, they account for only ~5% of the estimated heritability, suggesting that more loci await discovery.

We do not find association with any candidate genes we previously screened or any genes within the linkage region we previously reported (here the most significant discovery-stage signal within the linkage region was with rs17352829; $P = 9.83 \times 10^{-5}$)[13,14]. A rat model for MacTel carrying an insertion–deletion mutation within the gene Crb1 has been described[70]; however, we find no associated signal at this locus either.

Misdiagnoses can occur between MacTel and age-related macular degeneration (AMD). Hence, we searched for any overlap between our top association signals and those previously reported for AMD[71–73]. Among our suggestive association signals, we find the TIMP3 (TIMP metallopeptidase inhibitor 3)–SYN3 (synapsin III) locus at 22q12.3

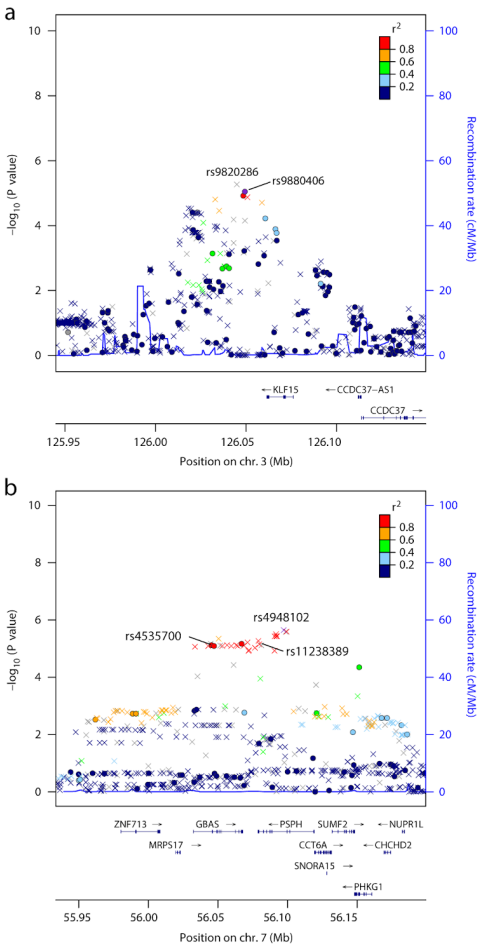


Figure 4 Regional plots of association. (a,b) Association results are shown for the analyzed SNPs, along with recombination rates (blue lines) and the location of known genes (labeled blue horizontal lines), within the two loci associated with suggestive significance at 3q21.3 (a) and 7p11.2 (b). For each plot, the x axis is scaled such that 100 kb flanks each end of the locus as defined by the first and last SNPs reaching suggestive significance ($P < 1 \times 10^{-5}$), and the y axis represents the $-\log_{10}$ (P value) of association for each SNP. Individual SNPs are represented as colored circles (genotyped) or crosses (imputed). For b, the sole purple cross represents the most significantly associated SNP at that locus in the discovery stage, and the other markers are colored according to their LD with that SNP; however, for a, the coloring is with respect to rs9880406 (the sole purple circle). Recombination rates and LD were estimated from the EUR 1000 Genomes Project March 2012 release, build hg19.

(Supplementary Table 2). TIMP3 is associated with both AMD and another maculopathy, Sorsby fundus dystrophy (SFD) (MIM 136900), which presents earlier in life than AMD but shares some clinical features[74]. The TIMP3 –SYN3 locus has also been suggestively associated with serine and glycine levels (see The Metabolomics GWAS Server[42,45]).

Major cell types of the retina include vascular cells, neuronal cells (rods and cones) and glial cells (microglia, Müller glia and astrocytes). Glial cells envelop the neurons and provide them with a homeostatic environment, supporting their survival and serving as an intermediary with blood vessels[75,76]. Müller glial cells remove neurotransmitters, including glutamate and glycine, from synaptic spaces to prevent eventual neurotoxicity and enable continued synaptic functioning. Müller cells also detoxify ammonia by expressing the enzyme glutamate–ammonia ligase (glutamine synthetase)[75,76]. Unlike retinal neuronal cells, healthy Müller cells are particularly resilient to stress, including anoxia, hypoglycemia and ischemia. Dysfunctional Müller cells could therefore lead to the pathological outcomes observed in MacTel. Indeed, we have previously reported abnormalities of Müller cells with MacTel[77,78] and observed MacTel characteristics following Müller cell ablation[79].

Retinal neurovascular coupling is important for the maintenance of local functionality, homeostasis and regulation of local blood supply[80]. The neurovascular unit that facilitates this function includes neurons, pericytes, endothelial cells and glia. Astrocytes and Müller glia regulate blood flow by vasoconstriction and vasodilation[81]; hence, aberrant functioning of these cells may affect this regulated control of vascular caliber and, thereby, blood flow[82]. Because we previously reported an association between MacTel and increased retinal vasculature calibers[82], our findings here linking MacTel risk alleles at 2q34 and 5q14.1 with increased vasodilation and with venular and arterial calibers, respectively (Table 3), may be relevant.

We found that the MacTel risk alleles at 1p12, 2q34 and suggestive locus 7p11.2 were previously associated with decreased serine levels (Table 3), and we find that serine and glycine serum levels are significantly lower in MacTel cases. The serine pathway involves ammonia and glutamate substrates, and it is plausible that a perturbation of this system due to defective enzyme activity might prevent Müller glia from performing their multifaceted functions normally. Altered expression of PHGDH , PSPH, ALDH1L1 or CPS1 (at 1p12, 7p11.2, 3q21.3 and 2q34, respectively) could lead to the observed decreases in serine levels and consequently to hyperammonemia and hyperglutamate conditions of a neurotoxic level, thereby causing retinal stress and damage. Müller glia and other cells react to retinal damage and stress by undergoing gliosis, a process that includes expression changes in glutamate–ammonia ligase and the induction of VEGFA expression. MEF2C (at 5q14.3, the strongest genome-wide associated locus) is under the control of VEGFA and may normally prevent stress-induced angiogenesis. Hence a defective variant of MEF2C may be unable to suppress angiogenesis during retinal damage.

Müller glia also remove neuron-derived carbon dioxide by expressing carbonic anhydrases (CAs)[75,76] that catalyze the reaction of carbon dioxide and water into hydrogen carbonate. One of the few drugs reported thus far with a positive effect on MacTel is acetazolamide, which reduces cystic lesions[83]. Acetazolamide is a CA inhibitor (Fig. 3). The hydrogen carbonate produced by Müller cells may be released into the vitreous body or bloodstream[76]. CPS1 uses hydrogen carbonate and ammonia as substrates for the urea cycle (chiefly in the liver).

Previously, we showed that glycolytic pathway genes have reduced expression in the MacTel macula in comparison to either non-macula retinal tissue of the same eye or healthy macula tissue[78] and a reduction in expression of glycolytic pathway genes following Müller cell ablation[79]. The former study observed a difference in expression of the enzyme phosphoglycerate kinase whose substrate is 3-phosphoglycerate. We find an association with PHGDH , which encodes an enzyme whose substrate is also 3-phosphoglycerate (Fig. 3), suggesting that both the glycolytic and serine pathways may be perturbed by PHGDH in the etiology of MacTel.

Despite MacTel and diabetes being associated (as reflected in our sample ascertainment; Table 1), our study finds very limited sharing of genetic risk at the three significant and two suggestive loci for the two disorders. Only the suggestive 3q21.3 locus has shown marginal association with clamp-based insulin sensitivity, whereas the other four loci have not been associated with any diabetes-related traits (Table 3).

Here we identified three validated and two suggestive loci with the maximal observed odds ratio at each locus ranging from 1.43 to 2.46 for each additional risk allele. Risk factors of similar, and smaller, effect size have been discovered using GWAS for AMD[72]. Further risk factors for MacTel will likely be identified using additional cohorts.

We report the first GWAS for MacTel and find genetic and metabolomic results implicating the glycine/serine metabolic pathway. We recognize that glycine and serine may not be the actual driver metabolites for MacTel. However, they may still be useful biomarkers for prediagnostic screening in at-risk individuals, and our findings may lead to metabolite or enzymatic supplementation as a course for prevention or slowing disease progression. Further studies are needed to validate these associations and to determine whether targeting those metabolites could become a useful strategy for screening or treatment. We also link MacTel with the glycolytic pathway and confirm a reported link between MacTel and retinal venular caliber, thereby increasing understanding of the disease.

URLs. PLINK, http://zzz.bwh.harvard.edu/plink/; PLINK1.9, https://www.cog-genomics.org/plink2; Genetic Power Calculator, http://zzz.bwh.harvard.edu/gpc/; LocusZoom, http://genome.sph.umich.edu/wiki/LocusZoom_Standalone; SHAPEIT, https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html; IMPUTE2, https://mathgen.stats.ox.ac.uk/impute/impute_v2.html; 1000 Genomes Project, http://www.1000genomes.org/; LINKDATAGEN, http://bioinf.wehi.edu.au/software/linkdatagen/; ExPASy, http://www.expasy.org/; Metabolomics GWAS Server, http://mips.helmholtz-muenchen.de/proj/GWAS/gwas/index.php; GTEx Project, http://gtexportal.org/; European Genome-phenome Archive (EGA), https://ega-archive.org/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge The Genomics Core Facility at the University of Utah School of Medicine for processing the Illumina Human Omni5 Exome BeadChips used in this study. This study was supported by the Lowy Medical Research Institute (La Jolla, California). We thank and acknowledge all of the participants of the MacTel Project (patients and controls) who have given their time, provided biological samples and undergone extensive testing for this study. The Age-Related Eye Disease Study (AREDS) control data set used for the discovery GWAS analyses described in this manuscript were obtained from the database at http://www.ncbi.nlm.nih.gov/ through database of Genotypes and Phenotypes (dbGaP) accession phs000429.v1.p1. Funding support for AREDS was provided by the National Eye Institute (N01-EY-0-2127). We would like to thank the AREDS participants and the AREDS Research Group for their valuable contribution to this research. We thank E. Agron for providing the AREDS diabetes summary statistics. We thank and acknowledge the staff and participants of the SABRE study who provided metabolomics data. Funding for the control group from the SABRE study was provided by the Wellcome Trust (WT082464), the British Heart Foundation (SP/07/001/23603) and Diabetes UK (13/0004774). Funding support for the control cohort at Columbia University was, in part, provided by NIH/NEI grant EY013435. This research was supported in part by the National Institute for Health Research (NIHR) Moorfields Biomedical Research Centre (London, UK). The views expressed are those of the authors and not necessarily those of the NIHR. This work was supported by Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS. M.B. is supported by an NHMRC Senior Research Fellowship (APP1002098) and an NHMRC Program Grant (APP1054618). Finally, we thank S. Freytag, T. Speed and G.K. Smyth for useful discussions with respect to statistical analyses and K. Khan for contributing serum samples for the metabolomics study.

AUTHOR CONTRIBUTIONS

T.S.S. designed the study, performed the analyses, interpreted the results, reviewed the literature on MacTel and wrote the manuscript. A.Q. performed the prediction modeling and other statistical analyses and helped write the manuscript. C.C. maintained the MacTel genetics database, including DNA isolation and preparation for genotyping from all MacTel subjects and controls and data organization. J.Z. performed TaqMan genotyping of the replication cohort and some data analysis. N.M. performed the GWAS SNP chip genotyping and helped write the manuscript. L.B. performed the GWAS SNP chip genotyping. L.S. organized patient databases and helped write the manuscript. R.B. performed the heritability and eQTL analyses and assisted with the metabolomics analysis. L.A.Y. obtained genetic material from the MacTel Project samples. M. Friedlander managed the study, helped interpret the results and helped write the manuscript. MacTel Project consortium members included clinicians and scientists who phenotyped the cohort of patients with MacTel and replication controls used in the study. C.A.E. and M. Fruttiger led the metabolomics study and interpreted the metabolomics data. M.L. led the genotyping group, helped design the study, interpreted the results and helped write the manuscript. R.A. led the genetics group, including obtaining genetic material from MacTel Project samples and the Columbia University controls and obtaining replication genotyping data, helped design the study, interpreted the results and helped write the manuscript. M.B. led the statistical analysis group, designed the study and helped write the manuscript.

COMPETING FINANICIAL INTERESTS

The authors declare no competing financial interests.

1. Gass, J.D.M. & Blodi, B.A. Idiopathic juxtafoveolar retinal telangiectasis—update of classification and follow-up-study. Ophthalmology 100 , 1536–1546 (1993).
2. Klein, R. et al. The prevalence of macular telangiectasia type 2 in the Beaver Dam Eye Study. Am. J. Ophthalmol. 150 , 55–62 (2010).
3. Aung, K.Z., Wickremasinghe, S.S., Makeyeva, G., Robman, L. & Guymer, R.H. The prevalence estimates of macular telangiectasia type 2: the Melbourne Collaborative Cohort Study. Retina 30 , 473–478 (2010).
4. Hannan, S.R., Madhusudhana, K.C., Rennie, C. & Lotery, A.J. Idiopathic juxtafoveolar retinal telangiectasis in monozygotic twins. Br. J. Ophthalmol. 91 , 1729–1730 (2007).
5. Chew, E.Y. Parafoveal telangiectasis and diabetic-retinopathy—reply. Arch. Ophthalmol. 104 , 972 (1986).
6. Clemons, T.E. et al. Medical characteristics of patients with macular telangiectasia type 2 (MacTel Type 2) MacTel project report no. 3. Ophthalmic Epidemiol. 20 , 109–113 (2013).
7. Menchini, U. et al. Bilateral juxtafoveolar telangiectasis in monozygotic twins. Am. J. Ophthalmol. 129 , 401–403 (2000).
8. Siddiqui, N. & Fekrat, S. Group 2A idiopathic juxtafoveolar retinal telangiectasia in monozygotic twins. Am. J. Ophthalmol. 139 , 568–570 (2005).
9. Gillies, M.C. et al. Familial asymptomatic macular telangiectasia type 2. Ophthalmology 116 , 2422–2429 (2009).
10. Hutton, W.L., Snyder, W.B., Fuller, D. & Vaiser, A. Focal parafoveal retinal telangiectasis. Arch. Ophthalmol. 96 , 1362–1367 (1978).
11. Oh, K.T. & Park, D.W. Bilateral juxtafoveal telangiectasis in a family. Retina 19 , 246–247 (1999).
12. Isaacs, T.W. & McAllister, I.L. Familial idiopathic juxtafoveolar retinal telangiectasis. Eye (Lond.) 10 , 639–642 (1996).
13. Parmalee, N.L. et al. Analysis of candidate genes for macular telangiectasia type 2. Mol. Vis. 16 , 2718–2726 (2010).
14. Parmalee, N.L. et al. Identification of a potential susceptibility locus for macular telangiectasia type 2. PLoS One 7 , e24268 (2012).
15. Ikram, M.K. et al. Four novel loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo. PLoS Genet. 6 , e1001184 (2010).
16. Sim, X. et al. Genetic loci for retinal arteriolar microcirculation. PLoS One 8 , e65804 (2013).
17. Brancati, F. et al. Autosomal dominant hereditary benign telangiectasia maps to the CMC1 locus for capillary malformation on chromosome 5q14. J. Med. Genet. 40 , 849–853 (2003).
18. Breugem, C.C. et al. A locus for hereditary capillary malformations mapped on chromosome 5q. Hum. Genet. 110 , 343–347 (2002).

19. Onishi, Y., Ohara, K., Shikada, Y. & Satomi, H. Hereditary benign telangiectasia: image analysis of hitherto unknown association with arteriovenous malformation. Br. J. Dermatol. 145 , 641–645 (2001).
20. Eerola, I. et al. Locus for susceptibility for familial capillary malformation ('port-wine stain') maps to 5q. Eur. J. Hum. Genet. 10 , 375–380 (2002).
21. Eerola, I. et al. Capillary malformation–arteriovenous malformation, a new clinical and genetic disorder caused by RASA1 mutations. Am. J. Hum. Genet. 73 , 1240–1249 (2003).
22. Revencu, N. et al. RASA1 mutations and associated phenotypes in 68 families with capillary malformation–arteriovenous malformation. Hum. Mutat. 34 , 1632–1641 (2013).
23. de Wijn, R.S. et al. Phenotypic variability in a family with capillary malformations caused by a mutation in the RASA1 gene. Eur. J. Med. Genet. 55 , 191–195 (2012).
24. Wooderchak-Donahue, W. et al. RASA1 analysis: clinical and molecular findings in a series of consecutive cases. Eur. J. Med. Genet. 55 , 91–95 (2012).
25. Henkemeyer, M. et al. Vascular system defects and neuronal apoptosis in mice lacking ras GTPase-activating protein. Nature 377 , 695–701 (1995).
26. Carr, C.W. et al. 5q14.3 neurocutaneous syndrome: a novel contiguous gene syndrome caused by simultaneous deletion of RASA1 and MEF2C . Am. J. Med. Genet. A. 155A , 1640–1645 (2011).
27. Lin, Q. et al. Requirement of the MADS-box transcription factor MEF2C for vascular development. Development 125 , 4565–4574 (1998).
28. Le Meur, N. et al. MEF2C haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. J. Med. Genet. 47 , 22–29 (2010).
29. Zweier, M. et al. Mutations in MEF2C from the 5q14.3q15 microdeletion syndrome region are a frequent cause of severe mental retardation and diminish MECP2 and CDKL5 expression. Hum. Mutat. 31 , 722–733 (2010).
30. Barbosa, A.C. et al. MEF2C, a transcription factor that facilitates learning and memory by negative regulation of synapse numbers and function. Proc. Natl. Acad. Sci. USA 105 , 9391–9396 (2008).
31. Lin, Q., Schwarz, J., Bucana, C. & Olson, E.N. Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. Science 276 , 1404–1407 (1997).
32. Bi, W., Drake, C.J. & Schwarz, J.J. The transcription factor MEF2C-null mouse exhibits complex vascular malformations and reduced cardiac expression of angiopoietin 1 and VEGF. Dev. Biol. 211 , 255–267 (1999).
33. Maiti, D., Xu, Z. & Duh, E.J. Vascular endothelial growth factor induces MEF2C and MEF2-dependent activity in endothelial cells. Invest. Ophthalmol. Vis. Sci. 49 , 3640–3648 (2008).
34. Xu, Z. et al. MEF2C ablation in endothelial cells reduces retinal vessel loss and suppresses pathologic retinal neovascularization in oxygen-induced retinopathy. Am. J. Pathol. 180 , 2548–2560 (2012).
35. Hao, H. et al. The transcription factor neural retina leucine zipper (NRL) controls photoreceptor-specific expression of myocyte enhancer factor Mef2c from an alternative promoter. J. Biol. Chem. 286 , 34893–34902 (2011).
36. Mears, A.J. et al. Nrl is required for rod photoreceptor development. Nat. Genet. 29 , 447–452 (2001).
37. Bessant, D.A. et al. A mutation in NRL is associated with autosomal dominant retinitis pigmentosa. Nat. Genet. 21 , 355–356 (1999).
38. Gelehrter, T.D. & Snodgrass, P.J. Lethal neonatal deficiency of carbamyl phosphate synthetase. N. Engl. J. Med. 290 , 430–433 (1974).
39. Demirkan, A. et al. Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. PLoS Genet. 11 , e1004835 (2015).
40. Raffler, J. et al. Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality. PLoS Genet. 11 , e1005487 (2015).
41. Sabater-Lleal, M. et al. Multiethnic meta-analysis of genome-wide association studies in > 100 000 subjects identifies 23 fibrinogen-associated loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. Circulation 128 , 1310–1324 (2013).
42. Shin, S.Y. et al. An atlas of genetic influences on human blood metabolites. Nat. Genet. 46 , 543–550 (2014).
43. Xie, W. et al. Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. Diabetes 62 , 2141–2150 (2013).
44. Mittelstrass, K. et al. Discovery of sexual dimorphisms in metabolic and genetic biomarkers. PLoS Genet. 7 , e1002215 (2011).
45. Suhre, K. et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature 477 , 54–60 (2011).
46. Raffler, J. et al. Identification and MS-assisted interpretation of genetically influenced NMR signals in human plasma. Genome Med. 5 , 13 (2013).
47. Illig, T. et al. A genome-wide perspective of genetic variation in human metabolism. Nat. Genet. 42 , 137–141 (2010).
48. Rhee, E.P. et al. A genome-wide association study of the human metabolome in a community-based cohort. Cell Metab. 18 , 130–143 (2013).
49. Yu, B. et al. Genetic determinants influencing human serum metabolome among African Americans. PLoS Genet. 10 , e1004212 (2014).
50. Danik, J.S. et al. Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: the Women's Genome Health Study. Circ Cardiovasc Genet 2 , 134–141 (2009).
51. Global Lipids Genetics Consortium et al. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45 , 1274–1283 (2013).
52. Kleber, M.E. et al. Genome-wide association study identifies 3 genomic loci significantly associated with serum levels of homoarginine: the AtheroRemo Consortium. Circ Cardiovasc Genet 6 , 505–513 (2013).
53. Kottgen, A. et al. New loci associated with kidney function and chronic kidney disease. Nat. Genet. 42 , 376–384 (2010).
54. Lange, L.A. et al. Genome-wide association study of homocysteine levels in Filipinos provides evidence for CPS1 in women and a stronger MTHFR effect in young adults. Hum. Mol. Genet. 19 , 2050–2058 (2010).
55. Pare, G. et al. Novel associations of CPS1 , MUT , NOX4, and DPEP1 with plasma homocysteine in a healthy population: a genome-wide evaluation of 13 974 participants in the Women's Genome Health Study. Circ Cardiovasc Genet 2 , 142–150 (2009).
56. Summar, M.L. et al. Relationship between carbamoyl-phosphate synthetase genotype and systemic vascular function. Hypertension 43 , 186–191 (2004).
57. van Meurs, J.B. et al. Common genetic loci influencing plasma homocysteine concentrations and their effect on risk of coronary artery disease. Am. J. Clin. Nutr. 98 , 668–676 (2013).
58. Williams, S.R. et al. Genome-wide meta-analysis of homocysteine and methionine metabolism identifies five one carbon metabolism loci and a novel association of ALDH1L1 with ischemic stroke. PLoS Genet. 10 , e1004214 (2014).
59. Klomp, L.W. et al. Molecular characterization of 3-phosphoglycerate dehydrogenase deficiency—a neurometabolic disorder associated with reduced l-serine biosynthesis. Am. J. Hum. Genet. 67 , 1389–1399 (2000).
60. Bouchard, L. et al. Mitochondrial 3-hydroxy-3-methylglutaryl-CoA synthase deficiency: clinical course and description of causal mutations in two patients. Pediatr. Res. 49 , 326–331 (2001).
61. Hong, M. et al. Isolation and characterization of cDNA clone for human liver 10-formyltetrahydrofolate dehydrogenase. Biochem. Mol. Biol. Int. 47 , 407–415 (1999).
62. Cahoy, J.D. et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J. Neurosci. 28 , 264–278 (2008).
63. Gariano, R.F. & Gardner, T.W. Retinal angiogenesis in development and disease. Nature 438 , 960–966 (2005).
64. Watanabe, T. & Raff, M.C. Retinal astrocytes are immigrants from the optic nerve. Nature 332 , 834–837 (1988).
65. Dorrell, M.I., Aguilar, E. & Friedlander, M. Retinal vascular development is mediated by endothelial filopodia, a preexisting astrocytic template and specific R-cadherin adhesion. Invest. Ophthalmol. Vis. Sci. 43 , 3500–3510 (2002).
66. De Laurenzi, V. et al. Sjögren–Larsson syndrome is caused by mutations in the fatty aldehyde dehydrogenase gene. Nat. Genet. 12 , 52–57 (1996).
67. Toomes, C. et al. Mutations in LRP5 or FZD4 underlie the common familial exudative vitreoretinopathy locus on chromosome 11q. Am. J. Hum. Genet. 74 , 721–730 (2004).
68. Veiga-da-Cunha, M. et al. Mutations responsible for 3-phosphoserine phosphatase deficiency. Eur. J. Hum. Genet. 12 , 163–166 (2004).
69. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45 , 580–585 (2013).
70. Zhao, M. et al. A new CRB1 rat mutation links Müller glial cells to retinal telangiectasia. J. Neurosci. 35 , 6093–6106 (2015).
71. Black, J.R. & Clark, S.J. Age-related macular degeneration: genome-wide association studies to translation. Genet. Med. 18 , 283–289 (2016).
72. Fritsche, L.G. et al. Seven new loci associated with age-related macular degeneration. Nat. Genet. 45 , 433–439 (2013).
73. Fritsche, L.G. et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat. Genet. 48 , 134–143 (2016).
74. Weber, B.H., Vogt, G., Pruett, R.C., Stohr, H. & Felbor, U. Mutations in the tissue inhibitor of metalloproteinases-3 (TIMP3) in patients with Sorsby's fundus dystrophy. Nat. Genet. 8, 352–356 (1994).
75. Giaume, C., Kirchhoff, F., Matute, C., Reichenbach, A. & Verkhratsky, A. Glia: the fulcrum of brain diseases. Cell Death Differ. 14 , 1324–1335 (2007).
76. Bringmann, A. et al. Müller cells in the healthy and diseased retina. Prog. Retin. Eye Res. 25 , 397–424 (2006).
77. Powner, M.B. et al. Perifoveal Müller cell depletion in a case of macular telangiectasia type 2. Ophthalmology 117 , 2407–2416 (2010).
78. Len, A.C. et al. Pilot application of iTRAQ to the retinal disease macular telangiectasia. J. Proteome Res. 11 , 537–553 (2012).
79. Chung, S.H. et al. Differential gene expression profiling after conditional Müller-cell ablation in a novel transgenic model. Invest. Ophthalmol. Vis. Sci. 54 , 2142–2152 (2013).
80. Usui, Y. et al. Neurovascular crosstalk between interneurons and capillaries is required for vision. J. Clin. Invest. 125 , 2335–2346 (2015).
81. Metea, M.R. & Newman, E.A. Glial cells dilate and constrict blood vessels: a mechanism of neurovascular coupling. J. Neurosci. 26 , 2862–2870 (2006).
82. Tikellis, G. et al. Retinal vascular caliber and macular telangiectasia type 2. Ophthalmology 116 , 319–323 (2009).
83. Chen, J.J. et al. Decreased macular thickness in nonproliferative macular telangiectasia type 2 with oral carbonic anhydrase inhibitors. Retina 34 , 1400–1406 (2014).
84. Artimo, P. et al. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. 40 , W597–W603 (2012).

ONLINE METHODS

Study population. Our MacTel Project consortium (Supplementary Table 1) recruited cases and controls at 23 participating clinical centers in seven countries (Australia, Germany, France, UK, Switzerland, Israel and United States). Informed written consent was obtained in accordance with ethics protocols for human subjects approved by the appropriate governing body at each site in accordance with the Declaration of Helsinki. Protocols and records of consent were centrally managed by the EMMES Corporation. The following ethics boards granted approval for human subject enrollment: Quinze-Vingts, Paris, France: Comité de Protection des Personnes Hôpital Saint-Antonie; Centre for Eye Research, Victoria, Australia: The Royal Victorian Eye and Ear Hospital; Clinique Ophtalmolgie de Creteil, Paris, France: Comité de Protection des Personnes Hôpital Saint-Antonie; Hospital Lariboisiere, Paris, France: Comité de Protection des Personnes Hôpital Saint-Antonie; Jules Stein Eye Institute, UCLA, California, USA: The UCLA Institutional Review Board; Lions Eye Institute, Nedlands, Australia: Sire Charles Gairdner Group Human Research Ethics Committee; Manhattan Eye, Ear and Throat Hospital, New York, USA: Lenox Hill Hospital Institutional Review Board; Moorfields Eye Hospital, London, UK: National Research Ethics Service; Retina Associates of Cleveland, Inc., Cleveland, Ohio, USA: Sterling Institutional Review Board; Save Sight Institute, Sydney, Australia: South Eastern Sydney Illawarra Area Health Service Human Research Ethics Committee–Northern Hospital Network; Scripps Research Institute, La Jolla, California, USA: Scripps Institutional Review Board; St. Franziskus Hospital, Munster, Germany: Ethik-Kommission der Arztekammer Westfalen-Lippe und der Medizinishchen Fakultat der Westfallschen Wilhelms-Universitat; The Goldschleger Eye Institute, Tel Hashomer, Israel: Ethics Committee The Chaim Sheba Medical Center; The New York Eye and Ear Infirmary, New York, USA: The Institutional Review Board of the New York Eye and Ear Infirmary; The Retina Group of Washington, Olympia, Washington, USA: Western Institutional Review Board; University of Bonn, Bonn, Germany: Rheinische Friedrich-Wilhelms-Universität Ethik-Kommission; University of Chicago, Chicago, Illinois, USA: The University of Chicago Division of Biological Sciences–The Pritzker School Institutional Review Board; University of Michigan, Ann Arbor, Michigan, USA: Medical School Institutional Review Board (IRBMED); University of Wisconsin, Madison, Wisconsin, USA: Office of Clinical Trials University of Wisconsin School of Medicine and Public Health; The Wilmer Eye Institute of Johns Hopkins University, Baltimore, Maryland, USA: Johns Hopkins School of Medicine Office of Human Subjects Research; Scheie Eye Institute University of Pennsylvania, Philadelphia, Pennsylvania, USA: University of Pennsylvania Office of Regulatory Affairs; University of Bern, Bern, Switzerland: Kantonale Ethikkommission Bern; John Moran Eye University of Utah, Salt Lake City, Utah, USA: The University of Utah Institutional Review Board; Bascom Palmer Eye Institute University of Miami, Miami, Florida, USA: The University of Miami Human Subjects Research Office; Columbia University, New York, New York, USA: Columbia University Medical Center Institutional Review Board Category 4 waiver for research involving specimens obtained from deidentified subjects. Participants were given a standardized ophthalmic examination, including best corrected visual acuity, fundus photography, fluorescein angiography, optical coherence tomography and blue light reflectance. Images were adjudicated at the Reading Center at Moorfields Eye Hospital, London. Diagnoses were made in accordance with the criteria described by Clemons et al.[85] on the basis of Gass and Blodi[1]. Retinal images were assessed for loss of transparency in the perifoveal region, dilated and telangiectatic blood vessels, especially in the temporal retina, and crystalline deposits. Participants were re-evaluated at regular intervals over the course of the study. MacTel Project consortium cases were used in both the GWAS discovery and replication stages.

GWAS discovery-stage control samples (N_{total} = 1,733) were predominantly sourced from the Age-Related Eye Disease Study (AREDS; $N_{subtotal}$ = 1,657) (refs. 86,87). The remainder came from the MacTel Project consortium ($N_{subtotal}$ = 76; most being spouses of MacTel cases) after confirmation as unaffected for MacTel by expert ophthalmologists. AREDS is a long-term natural history and clinical study of AMD and age-related cataracts. Patients were initially enrolled from 55–80 years of age. We used the control samples that, as described in dbGaP (phs000429.v1.p1), "are all Caucasian, do not have age-related macular degeneration (AMD) and were further screened to also exclude individuals with cataracts, retinitis pigmentosa, color blindness, other congenital eye problems, LASIK, artificial lenses, and other eye surgery."

Replication-stage control samples were of European ancestry and came from the Electronic Medical Records and Genomics (eMERGE) Cincinnati Children's Hospital Medical Center (CCHMC; $N_{subtotal}$ = 505) study and Columbia University ($N_{subtotal}$ = 629). The CCHMC controls were diagnosed with eosinophilic esophagitis, had an age range of 2–52 years and had Illumina Omni5 SNP chip data available from dbGaP (project "Better Outcomes for Children: GWAS from Cincinnati Children's Hospital Medical Center (CCHMC)–eMERGE Phase II data"; phs000494.v1.p1) (ref. 88). The Columbia University controls were confirmed as being of European ancestry by questionnaire during recruitment, have been used as controls for studies of AMD as described previously[72,89,90], were of an advanced age (mean age 74.8 ± 7.1 years), did not exhibit any distinguishing signs of macular or retinal disease after clinical examination by trained ophthalmologists and had no known family history of retinal disease.

DNA sample preparation and quality control. For MacTel Project consortium samples and Columbia University control samples, DNA was extracted from peripheral venous blood (Qiagen Blood maxi kit, 51194). DNA concentration was determined with the Qubit Fluorometer (Thermo Fisher Scientific). DNA samples of low purity were subjected to column purification (Qiagen Blood and Tissue kit, 69504).

GWAS genotyping and quality control. The GWAS discovery stage used a final quality-controlled data set of 2,209 samples (N_{cases} = 476; $N_{controls}$ = 1,733; Table 1).

AREDS genotypes were obtained from NCBI's dbGaP record (project "NEI Age-Related Eye Disease Study (AREDS)–Genetic Variation in Refractive Error Substudy"; phs000429, substudy of phs000001). AREDS samples were genotyped with Illumina Omni2.5 SNP chips. Genotypes for 2,182,680 variants/probes for 1,657 samples were downloaded from dbGaP, and those with minor allele frequencies (MAFs) <1%, or else monomorphic, were removed. Complementary biallelic (AT or CG) SNPs were removed because of strand uncertainty; these would otherwise cause problems when merging with the MacTel Project genotypes. For SNPs with identical chromosome identifiers and base-pair positions, only the one with the least missingness was kept. Identity-by-descent (IBD) sharing was assessed using PLINK1.9 (refs. 91,92), and a kinship coefficient threshold of 0.04 was determined to establish independence of samples. Further filtering took place as described below in conjunction with the MacTel Project samples.

From our MacTel Project cohort, a total of 704 samples (including MacTel cases and controls) were genotyped with the Illumina Omni5 SNP chip for the discovery stage. Illumina GenomeStudio software with default genotyping settings produced genotypes for a total of 4,641,218 variants/probes. DNA samples with genotype call rate <95% or unexpected or unusual sex-chromosome call rates were flagged and removed. Monomorphic probes or variants with MAF <1% were filtered out. Complementary biallelic (AT or CG) SNPs were removed because of strand uncertainty. For SNPs with identical chromosome identifiers and base-pair positions, only the one with the least missingness was kept. Sample independence was tested by assessing IBD sharing using PLINK1.9 (refs. 91,92) and applying a kinship coefficient threshold of 0.04. When IBD analysis suggested relatedness, a single sample was selected, which was usually the sample with the least missingness.

AREDS and MacTel project SNP genotypes were merged, and differential missingness was assessed between the two cohorts by Fisher's exact tests and an exclusion threshold of $P < 0.001$. A mock GWAS was performed on the 76 MacTel Project consortium controls against the 1,657 AREDS controls. A quantile–quantile plot of this analysis shows little deviation between the expected and observed P values (Supplementary Fig. 23). Further IBD analysis was performed to ensure no relatedness between AREDS and MacTel project samples. PCA of the samples was performed with EIGENSOFT and PLINK to identify any population substructure (Supplementary Figs. 2–4). This was achieved using 29,149 independent genome-wide SNPs selected with PLINK using the "--indep-pairwise" option; this assessed pairwise Pearson's correlations (r) within windows of 1,000 adjacent SNPs and greedily pruned

these SNPs until no such pairs remained with r^2 >0.04, before sliding the window a further 50 bp along.

Imputation of SNPs was performed with SHAPEIT (v2.r790) (ref. 93) and IMPUTE2 (v2.3.2) (ref. 94) following the software's best-practice guidelines. Reference genomes were obtained from the 1000 Genomes Project[95]. Compare Figure 1 with Supplementary Figure 24 to see the effect of imputation on the GWAS results.

Before statistical analysis, a final round of filtering removed any SNPs with >5% genotype missingness, Hardy–Weinberg equilibrium exact test $P < 1 \times 10^{-6}$ or MAF <5%.

GWAS statistical analysis. For the discovery stage, 2,209 samples (476 MacTel cases and 1,733 controls) with genotypes for 6,310,381 SNPs were analyzed by logistic regression with PLINK1.9 (ref. 92). PC1 from the PCA was added to the logistic regression model as a covariate in the discovery stage. The scree plot indicated that additional principal components would have a negligible contribution (Supplementary Fig. 3). Quantile–quantile plots (Supplementary Fig. 5) and Manhattan plots (Fig. 1 and Supplementary Fig. 24) were created in R with the qqman package[96]. The top associated loci ($P < 1 \times 10^{-5}$) are summarized in Supplementary Tables 2 and 3. Local association plots (Figs. 2 and 4) were constructed with LocusZoom[97]. The top SNP at each locus was also tested in PLINK using logistic regression with dominant and recessive modeling of the minor allele (A1) (Supplementary Table 5).

GWAS power calculations. We maximized the power of our GWAS discovery stage by using all MacTel Project cases available at the time. For GWAS of diseases such as MacTel, which are studied for the first time, it is not known what the expected effect size is. Power calculations were therefore performed post hocwith the Genetic Power Calculator[98], assuming a population prevalence of 0.001, $D' = 1$ (between the risk marker and true causal marker), $\alpha = 5 \times 10^{-8}$ (significance level), a range of genotypic relative risks and allele frequencies and using the allelic test (Supplementary Fig. 1). Given 476 cases and 1,733 controls, genotypic relative risks of 2 for genotype Aa and 4 for genotype AA, relative to the baseline genotype aa, and risk allele frequencies ranging from 0.20 to 0.60, would have >0.99 power. However, genotypic relative risks of 1.5 and 2.25 for the genotypes Aa and AA would only have a peak power of 0.53 with a risk allele frequency of ~0.4, thus demonstrating the limit of our GWAS discovery stage.

Replication analysis. The six loci that reached genome-wide significance were selected for replication (Fig. 2, Supplementary Figs. 25–27 and Supplementary Table 2). Three of these loci were subsequently ruled as false positives after attempts to technically validate them on an independent genotyping platform (TaqMan assays, Applied Biosystems) yielded uncorrelated genotypes (a low concordance rate). We identified two further loci on chromosomes 3 and 7 that showed suggestive evidence for association and warranted replication. This was based on their reported association with the serine and glycine metabolic pathways in other independent studies. Genotyping of the SNPs for replication at these five loci (three genome-wide significant and two suggestively associated) was performed by TaqMan assay in the new MacTel cases and Columbia University controls (Supplementary Fig. 28). Assays were purchased from Applied Biosystems as validated, inventoried SNP assays-on-demand or were submitted to the Applied Biosystems Assays-by-Design pipeline. The choice of these SNPs for replication was limited by the ability to create TaqMan assays. Furthermore, we ensured that at least one actual genotyped, rather than imputed, SNP was selected at each locus. The genotyping technique used was identical to that described previously[90]. Briefly, 5–10 ng of DNA was subjected to 50 cycles on an ABI 9700 384-well thermocycler, and plates were read on an Applied Biosystems 7900 HT Sequence Detection System. CCHMC control genotypes were generated with the Illumina Omni5 platform (data were unavailable for rs483180 and rs11238389).

Analysis in the replication cohort was performed with PLINK using a logistic regression framework and no covariates.

Power calculations were performed for the replication analysis. Given a replication cohort of 172 cases and 1,134 controls, a population prevalence of 0.001, $\alpha = 0.01$, $D' = 1$ (between the risk marker and true causal marker), and the odds ratios and allele frequencies of the top SNP at each locus, we estimated >97% power to replicate the top three genome-wide significant loci with the allelic test. At the two suggestively significant associated loci, 3q21.3 and 7p11.2, the power was 0.5943 and 0.7033, respectively.

Meta-analysis. Given the differences in ascertainment of the controls used for the discovery and replication stages and that a covariate was applied to the discovery-stage analysis but not to the replication stage, meta-analyses between the discovery and replication stages for the top hits were performed with a random-effects model in PLINK as previously described[99] (Table 2). Only two of the ten SNPs tested showed signs of differing effect sizes between the discovery and replication cohorts, while the results for the other eight SNPs were identical to modeling by fixed effects.

Heritability estimates. Heritability on the liability scale was estimated using a linear mixed model method, developed by Yang et al.[100] and adapted for dichotomous traits by Lee et al.[101], applied to the 1,093,805 directly genotyped SNPs from the discovery stage (no imputed SNPs). Applying a population prevalence of 0.0045% and 0.1% yielded h^2 of 0.21 and 0.74, respectively, with 5% of this h^2 explained by the three replicated and two suggestive loci (for both prevalence rates).

Prediction modeling. Logistic regression prediction modeling was performed in R with the GWAS discovery-stage cohort, using the strongest associated SNP that replicated at each of the five loci as predictors and with PC1 and sex as covariates (Supplementary Table 6). Initially, a variable selection procedure investigated the relevance of the interaction terms between sex and the SNPs and any epistatic effects. Once a stable model was selected, its predictive power was assessed. The genotypic effect for all markers was modeled as additive to be consistent with the discovery stage. We have shown that the additive model best describes the loci on chromosomes 1, 2, 5 and 7, and there is a marginal difference for the chromosome 3 loci between the additive and recessive models (Supplementary Table 5). To account for possible population stratification, we included PC1 using orthogonal polynomials of degree three (PC1_1, PC1_2 and PC1_3). Starting with the 'full' model, the variable selection step was performed using backward stepwise regression until convergence was achieved, as measured by the Bayesian information criterion (BIC). Fivefold cross-validation (CV) was used to perform variable selection with random subsets of the initial data, initialized multiple times with 100 different seeds. Blinding was used for case and control status for performance testing of the prediction model via an automated procedure in R. In total, 11 different models were observed from all 500 fitted models. One model predominated, being selected ~66% of the time, that included the main effects for all five SNPs (three significant and two suggestive), PC1 and the interaction term between sex and the SNP on chromosome 2 (column 4 in Supplementary Table 6). We did not observe any significant epistatic effect between SNPs. Using the R package ROCR[102], a ROC curve was plotted for this model with an AUC of 0.719 (and without PC1 an AUC of 0.707; Supplementary Fig. 13).

As PC1 was unavailable in the replication cohort (because of the lack of genotype data), this prediction model was applied to our replication cohort for validation after assessing the estimated effects of each SNP in the discovery cohort without PC1. We achieved an AUC of 0.679, thereby validating our results from the discovery cohort (Supplementary Fig. 14).

Sex-specific odds ratios at 2q34. At 2q34, we derived sex-specific odds ratios for the top most associated SNP (rs715) by performing logistic regression in R with orthogonal polynomials of degree three for PC1 (Supplementary Table 4). Considering the risk allele (which is the major allele), we found a borderline significant effect size in males (per-allele OR = 1.40, 95% CI = 1.08–1.81) and a significant association in females (per-allele OR = 2.58, 95% CI = 2.00–3.36).

Metabolomics analysis. Serum was collected from 50 MacTel cases and 50 controls from the wider MacTel Project cohort. Samples were well matched with regard to sex (25 females and 25 males in both cases and controls), age (average of 64 years in the MacTel cases and 63 years in the controls), diabetic status (38 non-diabetics and 12 type 2 diabetics in both cases and controls) and ancestry (48 Caucasian MacTel cases and 47 Caucasian controls).

All individuals fasted overnight, and blood was taken before noon. Around 5 ml of blood was collected in a clot-activating vacutainer tube (Vacutainer Plastic SST II Advance Tube with Gold Hemogard Closure, Becton Dickinson), left at room temperature for 30 min and then centrifuged for 5 min at 1,200*g*. The supernatant was collected, frozen and stored at −80 °C. Levels of 1,281 metabolites in the serum were measured by Metabolon. Briefly, this involved initial protein precipitation with methanol under vigorous shaking for 2 min (Glen Mills GenoGrinder 2000) followed by centrifugation. The resulting extract was analyzed by reverse-phase ultrahigh-performance liquid chromatography followed by tandem mass spectrometry (UPLC–MS/MS) with positive-ion-mode electrospray ionization (ESI). Peaks were quantified using the AUC technique. Quality control steps were applied to the 1,281 $\log_2$-transformed metabolite levels, thereby removing 403 for high missingness (>10% missing data for either MacTel cases or controls), 30 for high correlation (|Pearson's r| ≥ 0.95) between any pair of metabolites and 49 with abnormal distributions, thereby leaving a total of 799 metabolites for analysis. Missing metabolite data were then imputed with the minimum value for each metabolite across all samples. Quantile normalization of the data was performed using the R package limma[103]. Metabolite levels were then compared between the MacTel cases and controls by Student's t-test.

eQTL analysis. The 419 variants showing at least suggestive association ($P < 1 \times 10^{-5}$) with MacTel at the five loci were tested for association with *cis* gene expression levels in at least 40 tissues from 572 donors. Here *cis* is defined as being within ±1 Mb of the transcription start site of each gene. GTEx returns eQTL association results that are beyond specific gene- and tissue-derived significance thresholds.

Data availability. The MacTel consortium genotypes for 678 samples (with consent) of the 704 samples genotyped as part of the discovery stage and that support the findings of this study have been deposited in the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute and the CRG, under accession EGAS00001002249.

85. Clemons, T.E. et al. Baseline characteristics of participants in the natural history study of macular telangiectasia (MacTel) MacTel Project Report No. 2. Ophthalmic Epidemiol. 17, 66–73 (2010).
86. Bergeron-Sawitzke, J. et al. Multilocus analysis of age-related macular degeneration. Eur. J. Hum. Genet. 17, 1190–1199 (2009).
87. Age-Related Eye Disease Study Research Group. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E and beta carotene for age-related cataract and vision loss: AREDS report no. 9. Arch. Ophthalmol. 119, 1439–1452 (2001).
88. Kottyan, L.C. et al. Genome-wide association analysis of eosinophilic esophagitis provides insight into the tissue specificity of this allergic disease. Nat. Genet. 46, 895–900 (2014).
89. Gold, B. et al. Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. Nat. Genet. 38, 458–462 (2006).
90. Hageman, G.S. et al. A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. Proc. Natl. Acad. Sci. USA 102, 7227–7232 (2005).
91. Chang, C.C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015).
92. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
93. Delaneau, O., Marchini, J. & 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat. Commun. 5, 3934 (2014).
94. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).
95. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).
96. Turner, S.D. qqman: an R package for visualizing GWAS results using QQ and Manhattan plots. Preprint at bioRxiv https://doi.org/10.1101/005165 (2014).
97. Pruim, R.J. et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336–2337 (2010).
98. Purcell, S., Cherny, S.S. & Sham, P.C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 19, 149–150 (2003).
99. Borenstein, M., Hedges, L.V., Higgins, J.P. & Rothstein, H.R. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res. Synth. Methods 1, 97–111 (2010).
100. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569 (2010).
101. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. 88, 294–305 (2011).
102. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940–3941 (2005).
103. Ritchie, M.E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015).

Bibliography

1. J. Gass, Some problems in the diagnosis of macular diseases in *Symposium on Retinal Diseases*, (1977), pp. 268–270.

2. P. Charbel Issa, *et al.*, Macular telangiectasia type 2. *Prog. Retin. Eye Res.* **34**, 49–77 (2013/5).

3. J. D. Gass, B. A. Blodi, Idiopathic juxtafoveolar retinal telangiectasis. Update of classification and follow-up study. *Ophthalmology* **100**, 1536–1546 (1993).

4. R. P. Finger, *et al.*, Reading performance is reduced by parafoveal scotomas in patients with macular telangiectasia type 2. *Invest. Ophthalmol. Vis. Sci.* **50**, 1366–1370 (2009).

5. T. E. Clemons, *et al.*, The National Eye Institute Visual Function Questionnaire in the Macular Telangiectasia (MacTel) Project. *Invest. Ophthalmol. Vis. Sci.* **49**, 4340–4346 (2008).

6. E. L. Lamoureux, *et al.*, The longitudinal impact of macular telangiectasia (MacTel) type 2 on vision-related quality of life. *Invest. Ophthalmol. Vis. Sci.* **52**, 2520–2524 (2011).

7. R. Klein, *et al.*, The prevalence of macular telangiectasia type 2 in the Beaver Dam eye study. *Am. J. Ophthalmol.* **150**, 55–62.e2 (2010).

8. K. Z. Aung, S. S. Wickremasinghe, G. Makeyeva, L. Robman, R. H. Guymer, The prevalence estimates of macular telangiectasia type 2: the Melbourne Collaborative Cohort Study. *Retina* **30**, 473–478 (2010).

9. T. F. C. Heeren, F. G. Holz, P. C. Issa, "Macular Telangiectasia Type 2" in *Microperimetry and Multimodal Retinal Imaging*, (Springer, Berlin, Heidelberg, 2014), pp. 111–118.

10. T. E. Clemons, *et al.*, Baseline characteristics of participants in the natural history study of macular telangiectasia (MacTel) MacTel Project Report No. 2. *Ophthalmic Epidemiol.* **17**, 66–73 (2010).

11. T. E. Clemons, *et al.*, Medical characteristics of patients with macular

telangiectasia type 2 (MacTel Type 2) MacTel project report no. 3. *Ophthalmic Epidemiol.* **20**, 109–113 (2013).

12. D. Shukla, *et al.*, Type 2 idiopathic macular telangiectasia. *Retina* **32**, 265–274 (2012).

13. ,WikiJournal of Medicine/Medical gallery of Mikael Häggström 2014 - Wikiversity (August 22, 2017).

14. I. Leung, *et al.*, CHARACTERISTICS OF PIGMENTED LESIONS IN TYPE 2 IDIOPATHIC MACULAR TELANGIECTASIA. *Retina* (2017) https:/doi.org/10.1097/IAE.0000000000001842.

15. ,Anatomy Review. *Optical Coherence Tomography Scans* (August 23, 2017).

16. D. Pauleikhoff, *et al.*, Progression characteristics of ellipsoid zone loss in macular telangiectasia type 2. *Acta Ophthalmol.* **38**, S20 (2019).

17. R. Mathew, *et al.*, Agreement between time-domain and spectral-domain optical coherence tomography in the assessment of macular thickness in patients with idiopathic macular telangiectasia type 2. *Ophthalmologica* **230**, 144–150 (2013).

18. S. Müller, *et al.*, Contrast sensitivity and visual acuity under low light conditions in macular telangiectasia type 2. *Br. J. Ophthalmol.* (2018) https:/doi.org/10.1136/bjophthalmol-2017-311785.

19. P. C. Issa, F. G. Holz, "VEGF-Inhibition in Macular Telangiectasia Type 2" in *Anti-Angiogenic Therapy in Ophthalmology*, Essentials in Ophthalmology., (Springer, Cham, 2016), pp. 79–87.

20. E. H. Kupitz, T. F. C. Heeren, F. G. Holz, P. Charbel Issa, POOR LONG-TERM OUTCOME OF ANTI-VASCULAR ENDOTHELIAL GROWTH FACTOR THERAPY IN NONPROLIFERATIVE MACULAR TELANGIECTASIA TYPE 2. *Retina* **35**, 2619–2626 (2015).

21. E. Y. Chew, *et al.*, Ciliary neurotrophic factor for macular telangiectasia type 2: results from a phase 1 safety trial. *Am. J. Ophthalmol.* **159**, 659–666.e1 (2015).

22. E. Y. Chew, *et al.*, Effect of Ciliary Neurotrophic Factor on Retinal Neurodegeneration in Patients with Macular Telangiectasia Type 2: A Randomized Clinical Trial. *Ophthalmology* (2018) https:/doi.org/10.1016/j.ophtha.2018.09.041.

23. ,A phase III trial of NT-501 in patients with retinal telangiectasis type 2 - AdisInsight (August 17, 2017).

24. N. L. Parmalee, *et al.*, Identification of a potential susceptibility locus for macular telangiectasia type 2. *PLoS One* **7**, e24268 (2012).

25. N. L. Parmalee, *et al.*, Analysis of candidate genes for macular telangiectasia type 2. *Mol. Vis.* **16**, 2718–2726 (2010).

26. Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).

27. D. Houle, D. R. Govindaraju, S. Omholt, Phenomics: the next challenge. *Nat. Rev. Genet.* **11**, 855–866 (2010).

28. D. Gomez-Cabrero, *et al.*, Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* **8 Suppl 2**, I1 (2014).

29. 1000 Genomes Project Consortium, *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

30. ,Single nucleotide polymorphisms as genomic markers for high-throughput pharmacogenomic studies | Atlas of Science (October 10, 2018).

31. C. H. Johnson, J. Ivanisevic, G. Siuzdak, Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).

32. K. Bingol, Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods. *High Throughput* **7** (2018).

33. S. Z. Tan, P. Begley, G. Mullard, K. A. Hollywood, P. N. Bishop, Introduction to metabolomics and its applications in ophthalmology. *Eye* **30**, 773–783 (2016).

34. P. M. Visscher, *et al.*, 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

35. P. M. Visscher, M. A. Brown, M. I. McCarthy, J. Yang, Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

36. T. A. Manolio, *et al.*, Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

37. M. I. McCarthy, *et al.*, Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).

38. Y. Y. Teo, Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr. Opin. Lipidol.* **19**, 133–143 (2008).

39. T. W. Winkler, *et al.*, Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).

40. C. C. Laurie, *et al.*, Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).

41. S. Turner, *et al.*, Quality control procedures for genome-wide association

studies. *Curr. Protoc. Hum. Genet.* **Chapter 1**, Unit1.19 (2011).

42. International HapMap 3 Consortium, *et al.*, Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

43. J. J. Berg, *et al.*, Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **8** (2019).

44. J. J. Lee, *et al.*, Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

45. A. M. D. Livera, *et al.*, Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Anal. Chem.* **87**, 3606–3615 (2015).

46. A. M. De Livera, *et al.*, Normalizing and integrating metabolomics data. *Anal. Chem.* **84**, 10768–10776 (2012).

47. G. Blekherman, *et al.*, Bioinformatics tools for cancer metabolomics. *Metabolomics* **7**, 329–343 (2011).

48. A. Alonso, S. Marsal, A. Julià, Analytical methods in untargeted metabolomics: state of the art in 2015. *Front Bioeng Biotechnol* **3**, 23 (2015).

49. M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, D. Kim, Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16**, 85–97 (2015).

50. L. Tian, A. Quitadamo, F. Lin, X. Shi, Methods for population-based eQTL analysis in human genetics. *Tsinghua Sci. Technol.* **19**, 624–634 (2014).

51. A. Battle, C. D. Brown, B. E. Engelhardt, S. B. Montgomery, Genetic effects on gene expression across human tissues. *Nature Publishing Group* **550**, 204–213 (2017).

52. S.-Y. Shin, *et al.*, An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).

53. ,metabolomics gwas server (March 21, 2019).

54. G. Davey Smith, G. Hemani, Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–98 (2014).

55. G. Davey Smith, S. Ebrahim, "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).

56. D. M. Evans, G. Davey Smith, Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. *Annu. Rev. Genomics Hum.*

Genet. **16**, 327–350 (2015).

57. S. Greenland, J. Pearl, J. M. Robins, Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48 (1999).

58. W. G. Hill, T. F. C. Mackay, D. S. Falconer and Introduction to quantitative genetics. *Genetics* **167**, 1529–1536 (2004).

59. B. Bulik-Sullivan, *et al.*, An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

60. P. Turley, *et al.*, Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).

61. T. S. Scerri, *et al.*, Genome-wide analyses identify common variants associated with macular telangiectasia type 2. *Nat. Genet.* (2017) https:/doi.org/10.1038/ng.3799.

62. M. K. Ikram, *et al.*, Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo. *PLoS Genet.* **6**, e1001184 (2010).

63. X. Sim, *et al.*, Genetic loci for retinal arteriolar microcirculation. *PLoS One* **8**, e65804 (2013).

64. R. Madelaine, *et al.*, A screen for deeply conserved non-coding GWAS SNPs uncovers a MIR-9-2 functional mutation associated to retinal vasculature defects in human. *Nucleic Acids Res.* (2018) https:/doi.org/10.1093/nar/gky166.

65. W. Xie, *et al.*, Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. *Diabetes* **62**, 2141–2150 (2013).

66. K. Suhre, *et al.*, Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54–60 (2011).

67. J. Yang, *et al.*, Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

68. F. Chen, *et al.*, Methodological Considerations in Estimation of Phenotype Heritability Using Genome-Wide SNP Data, Illustrated by an Analysis of the Heritability of Height in a Large Sample of African Ancestry Adults. *PLoS One* **10**, e0131106 (2015).

69. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

70. O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1193–1198 (2012).

71. G. Gibson, Hints of hidden heritability in GWAS. *Nat. Genet.* **42**, 558–560 (2010).

72. S. H. Lee, N. R. Wray, M. E. Goddard, P. M. Visscher, Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).

73. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).

74. J. Friedman, T. Hastie, R. Tibshirani, glmnet: Lasso and elastic-net regularized generalized linear models. *R package version* **1** (2009).

75. ,GTEx Portal (May 23, 2019).

76. T. H. Pers, P. Timshel, J. N. Hirschhorn, SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418–420 (2015).

77. G. K. Smyth, "limma: Linear Models for Microarray Data" in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health., R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, S. Dudoit, Eds. (Springer New York, 2005), pp. 397–420.

78. J. L. Schafer, J. W. Graham, Missing data: our view of the state of the art. *Psychol. Methods* **7**, 147–177 (2002).

79. J. A. C. Sterne, *et al.*, Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338**, b2393 (2009).

80. S. van Buuren, K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," Faculty of Behavioural, Management and Social sciences (BMS). (2011) (March 10, 2017).

81. K. S. Kendler, M. C. Neale, Endophenotype: a conceptual analysis. *Mol. Psychiatry* **15**, 789–797 (2010).

82. I. I. Gottesman, T. D. Gould, The endophenotype concept in psychiatry: etymology and strategic intentions. *Am. J. Psychiatry* **160**, 636–645 (2003).

83. Marin L. Gantner, Kevin Eade, Martina Wallace, Michal K. Handzlik, Regis Fallon, Jennifer Trombley, Roberto Bonelli, et al, Serine and Lipid Metabolism in Macular Disease and Peripheral Neuropathy. *Accepted at New England Journal of Medicine.*

84. A. Penno, *et al.*, Hereditary sensory neuropathy type 1 is caused by the accumulation of two neurotoxic sphingolipids. *J. Biol. Chem.* **285**, 11178–11187 (2010).

85. M. Bertea, *et al.*, Deoxysphingoid bases as plasma markers in diabetes mellitus. *Lipids Health Dis.* **9**, 84 (2010).

86. T. Hornemann, Y. Wei, A. von Eckardstein, Is the mammalian serine palmitoyltransferase a high-molecular-mass complex? *Biochem. J* **405**, 157–164 (2007).

87. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting Linear Mixed-Effects Models using lme4. *arXiv [stat.CO]* (2014).

88. A. Kuznetsova, P. Brockhoff, R. Christensen, lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, Articles* **82**, 1–26 (2017).

89. I. Amelio, F. Cutruzzolá, A. Antonov, M. Agostini, G. Melino, Serine and glycine metabolism in cancer. *Trends Biochem. Sci.* **39**, 191–198 (2014).

90. M. Mehrmohamadi, X. Liu, A. A. Shestov, J. W. Locasale, Characterization of the usage of the serine metabolic network in human cancer. *Cell Rep.* **9**, 1507–1519 (2014).

91. C. F. Labuschagne, N. J. F. van den Broek, G. M. Mackay, K. H. Vousden, O. D. K. Maddocks, Serine, but not glycine, supports one-carbon metabolism and proliferation of cancer cells. *Cell Rep.* **7**, 1248–1258 (2014).

92. J. M. Dean, I. J. Lodhi, Structural and functional roles of ether lipids. *Protein Cell* **9**, 196–206 (2018).

93. S.-M. Park, *et al.*, 5q14.3 Microdeletions: A Contiguous Gene Syndrome with Capillary Malformation-Arteriovenous Malformation Syndrome and Neurologic Findings. *Pediatr. Dermatol.* **34**, 156–159 (2017).

94. X. R. Gao, H. Huang, H. Kim, Genome-wide association analyses identify 139 loci associated with macular thickness in the UK Biobank cohort. *Hum. Mol. Genet.* (2018) https:/doi.org/10.1093/hmg/ddy422.

95. P. J. Spring, *et al.*, Autosomal dominant hereditary sensory neuropathy with chronic cough and gastro-oesophageal reflux: clinical features in two families linked to chromosome 3p22--p24. *Brain* **128**, 2797–2810 (2005).

96. P. L. Tan, C. Bowes Rickman, N. Katsanis, AMD and the alternative complement pathway: genetics and functional implications. *Hum. Genomics* **10**, 23 (2016).

97. P. S. Graham, *et al.*, Genome-wide association studies for diabetic macular edema and proliferative diabetic retinopathy. *BMC Med. Genet.* **19**, 71 (2018).

98. Y. Kihara, *et al.*, Estimating Retinal Sensitivity Using Optical Coherence Tomography With Deep-Learning Algorithms in Macular Telangiectasia

Type 2. *JAMA Netw Open* **2**, e188029 (2019).

99. R. Ratnapriya, *et al.*, Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat. Genet.* (2019) https:/doi.org/10.1038/s41588-019-0351-9.